1-1-2012

Novel Sequence-Based Method for Identifying Transcription Factor Binding Sites in Prokaryotic Genomes

Gurmukh Sahota
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Gary Stormo, Chair
Jeremy Buhler
Justin Fay
Jeffrey Gordon
James Havranek
David Wang

Novel Sequence-Based Method for Identifying Transcription Factor Binding Sites in Prokaryotic Genomes

by

Gurmukh Singh Sahota

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2012

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes

by

Gurmukh Singh Sahota

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2012

Professor Gary D. Stormo, Chairperson

Computational techniques for microbial genomic sequence analysis are becoming increasingly important. With next-generation sequencing technology and the human microbiome project underway, current sequencing capacity is significantly greater than the speed at which organisms of interest can be experimentally probed. We have developed a method that will primarily use available sequence data in order to determine prokaryotic transcription factor binding specificities.

The prototypical prokaryotic transcription factor (TF) contains a helix-turn-helix (HTH) fold and bind DNA as homodimers, leading to their palindromic motif specificities. The connection between the TF and its promoter is based on the autoregulation phenomenon noticed in *E. coli.* Approximately 55% of the TFs analyzed were estimated to be autoregulated. Our preliminary analysis using RegulonDB indicates that this value increases to 79% if one considers the neighboring operons.

Given the TF family of interest, it is necessary to find the relevant TF proteins and their associated genomes. Due to the scale-free network topology of prokaryotic systems, many of the transcriptional regulators regulate only one or a few operons. Within a single genome, there would not be enough sequence-based signal to determine the binding site using standard computational methods. Therefore, multiple bacterial genomes are used to overcome this lack of signal within a single genome.

We use a distance-based criteria to define the operon boundaries and their respective promoters. Several TF-DNA crystal structures are then used to determine the residues that interact with the DNA. These key residues are the basis for the TF comparison metric; the assumption being that similar residues should impart similar DNA binding specificities. After defining the sets of TF clusters using this metric, their respective promoters are used as input to a motif finding procedure.

This method has currently been tested on the LacI and TetR TF families with successful results. On external validation sets, the specificity of prediction is ~80%. These results are important in developing methods to define the DNA binding preferences of the TF protein residues, known as the "recognition code". This "recognition code" would allow computational design and prediction of novel DNA-binding specificities, enabling protein-engineering and synthetic biology applications.

Acknowledgements

To simply acknowledge the people who are listed below would be an understatement of the deep gratitude I feel in their contribution to the process of completing a PhD. The majority of the funding that helped me completed this project was provided by the National Institutes of Health through two training grants, the computational biology training grant (GM008802) and the MSTP training grant (GM07200).

For providing me with this opportunity in such a rich academic environment, I would like to thank several separate entities at Washington University in St. Louis. To the medical school deans, Dr. Edwin Dodson, Dr. Koong-Nah Chung, Dr. Leslie Kahl and Dr. Alison Whelan, I thank you for giving me the opportunity to train at such an excellent medical school while being so supportive of me as an individual. To the graduate school and the specifically, the program formerly known as the Computational Biology program, the opportunity to learn from the leading experts in my PhD field. And finally to the MSTP with Liz Bayer, Christy Durbin, Linda Perniciaro, Andrew Richards, Dr. Brian Sullivan, headed by initially Dr. Dan Goldberg and now Dr. Wayne Yokoyama. I would like to thank you for shepherding me through the process of an MD/PhD and always making sure that I had my i's dotted and my t's crossed.

I would like to thank my thesis committee chaired by Dr. Jeff Gordon and comprised of Dr. Jeremy Buhler, Dr. Justin Fay, Dr. James Havranek and Dr. David Wang. They have provided me with excellent scientific feedback and support. They were not only nurtured the seedling of this thesis, but also my growth as a scientist.

This growth would also not have been possible without the support of the Stormo lab. All of the members were always ready to give critical feedback, but I would like to specifically mention 3, Aaron Spivak, Yue Zhao and Ryan Christensen. Aaron, I thank for helping me translate my "theoretical" molecular biology knowledge into practical problem solving in some of my many projects in the lab. Yue always made sure that I could defend what I was saying scientifically (no hand-waving allowed) and held my feet to the proverbial fire. But he was also always available to talk about the current hurdle I was trying to overcome. And finally to Ryan, for always being willing to talk be it about science, life or otherwise. He showed me how to successfully share intellectual credit as part of a team and was always willing to think through ideas with me. Also, he had the uncanny ability to have some code or program that I needed and was always more than willing to share it.

Last but not least, I thank my PI and mentor, Dr. Gary Stormo. Gary taught me a lot of things, some of which can be found in scientific texts and archives and others which cannot. In providing an environment where there was a lot of intellectual freedom, he supported me to pursue many tangential projects that piqued my scientific curiosity, but in the end made sure that I kept my eye on the prize. Through example, he showed me how to be an ethical and successful scientist. For these lessons and the opportunity to learn how to conduct independent science, I am grateful.

Finally, I would like to thank my family. Without their love and support, I would not be here today. There are two members who I would like to thank specifically. The first is my sister Dr. Sukhvinder Nagi. She inspired me to follow my heart and introduced me to lab research by example. She set me on the path of research and science. The

second is my wife, Dr. Puneet Sahota; to her I owe the debt of gratitude of making sure that I finished what I had started. She has been my companion along this path and made sure that I was well supported during both the easy times and the hard times. Finally, I mention my son, Anand Sahota. He kept me awake for some of the nights, during which I had important breakthroughs on this PhD project as well.

Table of Contents

List of Tables

List of Figures

Figure A2.S14. *Batf* $^{-/-}$ T cell do not protect against EAE.

Chapter 1: Background

Introduction

There has been increasing interest in understanding microbes both in the context of their symbotic relationship with their human hosts, as in the microbiome project (Friedrich, 2008; Peterson et al., 2009; Turnbaugh et al., 2007), as well as in the context of disease as exemplified by the tuberculosis projects (Cole et al., 1998; Schürch et al., 2010). This interest has lead to an increasing number of sequenced microbial genomes. Much of the current computational analysis of this sequence information focuses on functional classification of genes or metabolic network reconstruction. As of yet, there has been little large-scale computational work focused on understanding transcriptional regulation in these prokaryotic systems.

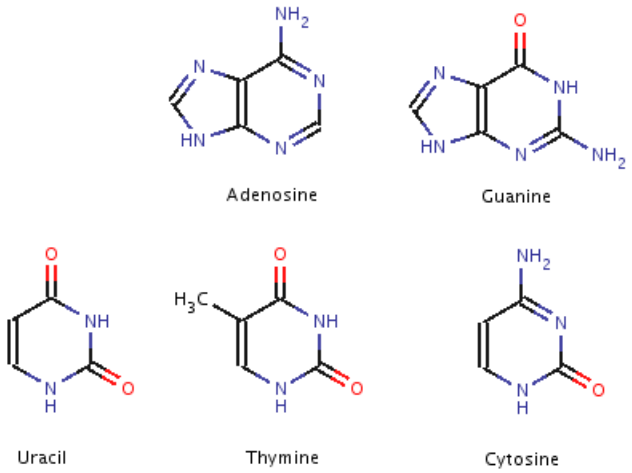
Unraveling the gene regulation network is an important step in a deeper understanding of the functional complexity of an organism. Gene regulation is a critical in the adaptation of the cell to its environment via selective tuning of protein levels. Transcription factors are proteins that are able to bind DNA in order to transcribe the DNA into RNA. In prokaryotic systems, the transcription factors (TF) are the main mechanism of selective gene regulation via cognate binding sites in the promoters of regulated genes. This thesis describes a method that combines the increasing amount of sequence data with a few simplifying assumptions in order to generate large-scale predictions of the binding sites of prokaryotic transcription factors.

Transcriptional Regulation

The central dogma of molecular biology states that information flows from DNA to RNA to proteins. DNA is composed of 4 nucleotide bases, adenine (A), guanine (C),

cytosine (G) and thymine (T). RNA replaces the thymine (T) with uracil (U) and has a 2' OH as well, leading to increased lability.

Figure 1.1. Nucleotide bases.

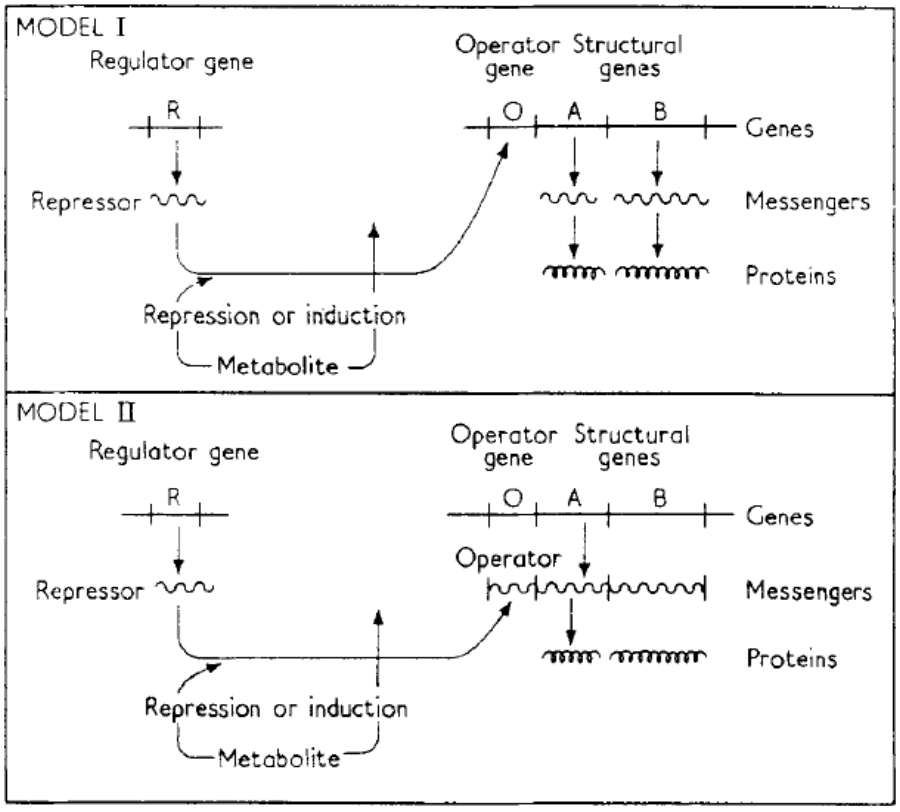


In this context, RNA was commonly thought of as a temporary intermediary between DNA and protein, however there has been a growing body of literature discussing functional RNA enzymes, ribozymes and riboswitches (Guerrier-Takada et al., 1983; Kruger et al., 1982; Lai, 2003; Serganov, 2009). The process by which information flows from DNA to RNA is called transcription. Translation is the process of converting the information encoded in the RNA into proteins. The DNA that eventually encodes a functional product is generally referred to as a gene. Regulation of gene expression levels is mainly controlled during the process of transcription. Both transcription and translation occur in three distinct phases: initiation, elongation, and termination. In the context of transcriptional regulation, the main control point is the during the initiation phase.

The canonical example of transcriptional control in prokaryotes was developed from the study of the lac operon nearly a half century ago (Jacob and Monod, 1961). In their study of the lac operon, they discovered the difference between the regulator (TF), a

trans-acting element, and the operator (TFBS), a *cis*-acting element. They hypothesized that these two separate components were required in order to activate or repress the "gene" of interest. In addition, they hypothesized a model, Model II, of coordinate expression of a set of genes, that they named an "operon" as shown in Figure 1.2.

Figure 1.2. Models of regulation of protein synthesis. Reproduced with permission from Journal of Molecular Biology (JMB).



The regulatory region immediately upstream of the operon is called the promoter, that contains the regulatory elements necessary for transcription initiation. The result of this archtecture is that these sets of genes in the operon are both co-transcribed as well as co-regulated. This co-regulation and co-transcription is usually because the operons encode genes that are part of a common pathway or are members of a complex. Many times these operons include their own regulators, up to 55% of *Escherichia coli* operons

are autoregulatory. A simple analysis of RegulonDB 6.7 (Gama-Castro et al., 2008), a database based on *E. coli*, increases this value to 79% if one includes the promoters of adjacent upstream and downstream operons.

DNA-dependent RNA polymerase is the main enzyme responsible for transcribing the DNA into RNA (Ebright, 2000). The complete holoenzyme consists of six subunits: α2ββ'σω. The two α subunits assemble the enzyme and bind regulatory factors. The β subunit catalyzes the synthesis of RNA, both chain initiation and elongation. Nonspecific DNA binding is mediated via the β' subunit. The function of the ω subunit is not entirely clear, but currently it is believed to help with the folding of the β' subunit and assembly of the RNA polymerase (Mathew and Chatterji, 2006). The final subunit, σ, targets the RNA polymerase to certain promoters while decreasing the overall nonspecific promoter binding.

Promoter architecture

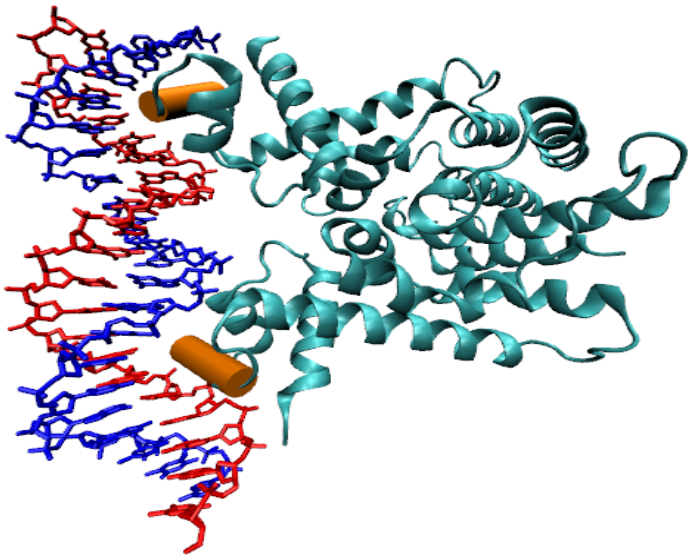
The RNA polymerase holoenzyme interacts with several specific cis-DNA elements that together form the "core" promoter. These elements include the transcription start site (TSS), -10 element, -35 element, and the UP element (centered at -50). The -10 element, also known as the Pribnow box (Pribnow, 1975), serves a similar function to the TATA box in eukaryotes, but is essential for initiation of transcription and has a consensus sequence of TATAAT. The -35 element is a heptamer with a consensus sequence of TTGACAT and typically spaced about 17 +/- base pairs upstream of the -10 element. There do not appear to be any bacterial promoters with both the -10 and -35 consensus sequences, perhaps because that would create too tight of a bond between the RNA polymerase and the DNA and actually inhibit its translocation. The UP element and

the -35 element seem to be interchangable such that, in the presence of an UP element, the -35 element is not required for transcription (Estrem et al., 1998). These sets of elements in the core promoter localize the initial binding of the RNA polymerase. After binding to the DNA, the polymerase converts from the closed complex to the open complex. Eventually for the polymerase to continue the elongation process, the σ factor must be removed. In addition to this core promoter, the dynamic control of transcription in prokaryotes is mainly accomplished via proteins that bind specifically to the upstream promoter DNA regions, namely transcription factors. Through interactions with RNA polymerase, their binding strength to the DNA is has a correlated effect to the amount of RNA that is eventually produced and thus the protein product.

Transcription factors are able to bind specifically to DNA *cis*-elements modulating their binding in response to external stimuli, generally via protein or metabolite allosteric interactions. In making this connection, they becoming the critical link in transferring information from the environment into functional consequences. The prototypical prokaryotic transcription factors are mainly of the helix-turn-helix (HTH) fold family, and selectively bind to the major groove of DNA using the second helix of the TF. Approximately 80% of *E. coli* TFs are members of the HTH superfamily. These HTH TFs generally bind as homo-dimers (and more rarely as homo-tetramers) leading to a palindromic binding specificity, however in certain families they can bind direct repeats as well.

Figure 1.3. HTH binding model. The structure shown is adapted from PDB (2yvh). The structural elements highlighted in orange are the recognition helices of the HTH binding domain. As can be seen, they fit into the major groove of the DNA where they can make

base-specific contacts. Figure created using VMD (Humphrey et al., 1996).



Horizontal gene transfer and phylogeny

Horizontal gene transfer is a mechanism by which different bacteria can share genetic material. This process is believed to be important in the rapid emergence of drug resistance across multiple different bacterial species given a few resistant bacteria (Hawkey and Jones, 2009). Given a local regulatory structure, the sharing of genetic material between different bacterial species via horizontal gene transfer would be more effective due to a lower risk of importing of an entirely unregulated set of operons. This process of horizontal gene transfer also creates confusion within a phylogenetic analysis, which generally depend on sequence differences between taxa in order to generate an evolutionary tree. Phylogenetic approaches are usually based on a molecular clock hypothesis, where changes in the DNA are an evolutionary process that is roughly

constant over a unit of time and these eventual processes lead to speciation and adaptation. With lateral transfer, the problem is that changes in the DNA are no longer only due to top-down speciation events over time, from parent to child. This complication becomes important in the context of selecting methods in order to analyze multiple microbial genomes. Many advanced motif finding approaches rely on orthology and phylogeny in order to more effectively determine putative motifs. In addition, as will be discussed later, the assumption of sequence independence between phylogenetically distant relatives may not apply in the context of horizontal transfer, thus a normalization procedure based on phylogeny may lead to spurious signals.

Experimental determination of binding sites.

Many experimental methods have been used in order to identify interactions between transcription factors (TF) and their corresponding DNA binding elements. In antibody-based TF detection assays, one can either directly detect the TF or tag every TF with an epitope such as a GST tag. The benefit of a direct TF antibody is that it will detect a native form of the TF and can be easily multiplexed, although if the TF is modified from the form used to generate the antibodies, such as by post-translational modification, the antibodies may cross-react or not bind at all. Additionally, direct detection requires generating antibodies to each TF of interest. The epitope tag can alleviate some of these issues by inserting a constant region that can be is detected by an antibody regardless of the linked TF. However, the tag can potentially impede or change native TF-DNA interactions, and thus needs to be used with care.

Electrophoretic mobility shift assay (EMSA) uses gel shifting in order to determine DNA binding. Specifically, a TF is purified and labeled DNA oligmer probes

are mixed and introduced into a polyacrylimide gel. An electric voltage is introduced across this gel and the protein-DNA mixture, stabilized in part by the gel-matrix moves via electrophoresis. Then the complex is visualized using florescence, UV or another appropriate method depending on the labeling strategy of the DNA. The TF-DNA complex will be at a larger molecular weight than the labeled oligimer probe. If the starting concentrations of the TF and the oligmer probe is known, then the affinity can be estimated from the relative intensities of the bound and unbound fractions. If the TF concentration is not known, then one can titrate the labeled probe and measure against a known standard (Buratowski and Chodosh, 2001). If the protein is not purified, it is also possible to use an antibody specific for the TF of interest generating a "supershift".

Systematic evolution of ligands by exponential enrichment (SELEX) (Tuerk and Gold, 1990) is an *in-vitro* method that is used to identify target sequences that have affinity for the TF of interest. The TF is affixed to a substrate and an input library of potential DNA oligmers is flowed over the substrate. Iterative rounds of selection and elution followed by amplification are performed. At the end of several rounds, the strongest binders are most likely to remain. Recently, a method has been developed so that only a single round of SELEX is required to determine the specificity of a transcription factor using a maximum-likelihood model and the background frequencies of the input library (Zhao et al., 2009).

In protein-binding microarrays (PBMs) (Berger et al., 2006; Berger and Bulyk, 2009), microarrays that cover all of the sequence space of a certain length are generated, using a linearized de Bruijn sequence. TFs are flowed over this microarray and binding associations are detected using an antibody to the TF. A control experiment to detect

cross-reaction of the microarray and the antibody is also performed. The ratio between the control and sample are used to calculate an enrichment score for each potential DNA binding sequence.

In the genomic context, one can implement protocols based on chromatin immunoprecipitation (ChIP). There are two major variants that are used to determine TF-DNA interactions across a genome, ChIP-chip (Buck and Lieb, 2004) and ChIP-Seq (Johnson et al., 2007). The main difference between the two is the readout method, the former uses a microarray and the latter uses sequencing. The overall protocol consists of taking cells that overexpress an epitope-tagged TF or unmodified cell lines, crosslinking the DNA to the DNA-binding proteins, sheering the genomic DNA, purifying the TF-DNA crosslinked complex of interest (via immunoprecipitation with an anti-TF antibody or an anti-epitope antibody), un-crosslinking the DNA and then use one of the two readout methods to determine the sequence of the DNA that was bound by the TF during the crosslinking process. In the ChIP-Seq protocol, the same rapid sequencing technologies that have increased the amount of bacterial genomic sequence, are also being applied to understand TF-DNA interactions. However, to assess all of the transcription factors in a genome, the protocol would require antibody pulldowns of every TF of interest, either in a successive manner by epitope tagging every TF, or by creating antibodies to every TF of interest in the native form.

The bacterial one-hybrid (B1H) has also been used to determine the specificity of DNA interactions. The method relies on two plasmids, one containing the TF fused to the ω subunit of RNA polymerase and the second containing a randomized binding region (Meng et al., 2005). A major advantage of the B1H system is that the protein does not

need to be purified, however, given that this is a bacterial system, it is likely that some bacterial TF's would cross-react with the transcriptional machinery of the *E.* coli host and cause errors in interpretation.

Motif Models

Representation of a binding site can take multiple forms. Most simply, one could list the series of sequences that are bound. There is no loss of information in the enumerative model, however there is also no generalization or easy ability to interpolate missing data points. A summarization of the most likely to be bound set of sequences could be a "consensus" sequence. The consensus sequence is generally described using the IUPAC degenerate alphabet or a regular expression. The IUPAC degenerate alphabet consists of 15 characters that describe all possible combinations of nucleotides, A, C, G and T. For example, S, for strong hydrogen bond, represents C or G, whereas N represents any nucleotide A, C, G or T. Again, the issue is that the consensus sequence may not reflect the relative influence of a nucleotide at each position on the overall binding probability (for example, does an S imply equiprobably C/G or was C more favorable at that position). A more numerical approach would solve this problem of relative ranking of binding sites. The position weight matrix (PWM) or position specific scoring matrix (PSSM) is such a method of describing the motif. In this model, each column, generally representing a single nucleotide position, is independent and the net energy is additive. There are multiple methods to construct a PWM given a set of sequences, but generally one counts the number of nucleotides at each position and divides that count by the sum over all nucleotides at that column creating a position frequency matrix (PFM). Many times, a small pseudocount is added to each nucleotide

count in order to mitigate sampling errors, and numerical complications of taking the logarithm of zero. This PFM is transformed into a PWM by taking the log-likelihood ratio of the observed frequency of a nucleotide at that position vs the expected frequency, based on the background nucleotide distribution. The benefit of a PWM is that any potential sequence can be scored whether it was observed or not. A common convention is that lower scoring sites are correlated with higher energy of binding and higher scoring sites are correlated with lower energy of binding. Since each column is independent, the more conserved columns can often contribute more to the overall energy of binding, which usually means they are more important in determining the strength of the TF-DNA interaction (G D Stormo, 2000). The assumption of independence between adjacent nucleotide positions has been challenged, but in most cases, it appears to be a relatively robust (Benos et al., 2002).

Motif Finding

The prior section assumed that the sites that contributed to the binding were known and all that was needed was a model to describe their interaction with the DNA. The set of transcription factors binding sites is rarely known *a-priori*, so both the model and the set of sites are being learned simultaneously. The problem is that given a set of sequences that are likely to contain a binding site, find a set of statistically overrepresented words contained within, or a statistically significant alignment of possible binding sites. These sets of words may be the binding sites for the transcription factor of interest.

Under the assumption of a fixed width motif of length l, simplest method of accomplishing this task would be to count the occurrences of all the l-long words, lmers,

within the set of sequences. The problem with this simplistic method is that it requires all positions to be exactly identical. Many TFs bind a set of similar sequences rather than imposing the restriction that all sequences be completely identical. A variant of this method leads to the classical computer science problem of the planted l,d motif (Pevzner and Sze, 2000). In this formulation, the group of the l long words with up to d differences is considered to be part of the same motif. These two methods generally enumerate all possible motifs and select the best representative motif from that set.

Alternatively, one can use methods that are based on iteratively building the PWM as opposed to enumerating all possible lmers. There are three major components to all motif finders: a motif model, an objective function and a search algorithm. The motif model is generally a PWM. The objective function is a measure of the likelihood of that specific PWM being enriched in the set of sequences rather than simply a null model. There are three main algorithms used in the search step, including greedy algorithms, expectation-maximization (EM), and a variation of EM called Gibbs sampling. Representing these three classes are CONSENSUS (Hertz and Stormo, 1999; Stormo and Hartzell, 1989), MEME (Bailey and Elkan, 1994) and Gibbs sampler (Neuwald et al., 1995) respectively.

The CONSENSUS algorithm creates a set of potential start matrices using all possible lmers. In each subsequent iteration, the best matching lmer from the sequence is added to the current set of matrices, and the top scoring PWMs are retained for the next round. This process is continued for all sequences. This algorithm is greedy, because the order of the input sequences will matter to the final result of the algorithm. An alternative method is to use EM. The EM algorithm is commonly used to generate the maximium

likelihood estimate of a set of parameters via alternating cycles of expectation of a likelihood function using the current parameters followed by maximization of the parameters based on the current expectation of the likelihood function. In this regards, EM can be viewed as a local descent algorithm. MEME uses EM with a series of comprehensive start points in order to generate a set of likely motifs. Gibbs sampling is very similar to EM, with the exception that instead of a simple maximization step, Gibbs uses a roulette wheel selection to decide the next set of parameter values. This allows the algorithm to potentially make a move that is less favorable, thus allowing it to climb out of local minima. Incorporating additional data lead to newer generations of motif finders that have attempted to capitalize on the idea that orthologous transcription factors should bind similar sequences in orthologous promoter regions (Prakash et al., 2004; Saurabh Sinha et al., 2004; Siddharthan et al., 2005; Wang and Stormo, 2003).

Overview of this thesis

This thesis describes a method that can be applied to a series of bacterial genomes in order to determine the transcription factor binding sites without additional experimental data. Chapter 2 describes the overall approach and its ability to determine the binding specificities of two HTH families, namely LacI and TetR. The results of Chapter 2 showed some limitations in MEME in its application to this specific problem, that spurred the development of an improved motif finding algorithm, including gaps and multiple orientations of binding as discussed in Chapter 3. Finally, future directions for this project are described in Chapter 4. Two additional appendices are attached which describe ancillary projects related to motif finding in non-prokaryotic systems that were performed before this thesis project was undertaken. These ancillary projects provided

insights as well as a set of tools and codebase that helped in accomplishing this thesis. Specifically, an alternative gibbs-sampling based gapped motif finder, much of the SVG logo generation framework, and many of the motif input/output and analysis routines were implemented in the scope of these additional projects.

References

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.

Benos,P.V. et al. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, **30**, 4442-4451.

Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protocols*, **4**, 393-411.

Berger,M.F. et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol*, **24**, 1429-1435.

Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349-360.

Buratowski,S. and Chodosh,L.A. (2001) In, *Mobility Shift DNA-Binding Assay Using Gel Electrophoresis*, Current Protocols in Molecular Biology. John Wiley & Sons, Inc., pp. 12.2.1-12.2.11.

Cole,S.T. et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537-544.

Ebright,R.H. (2000) RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J. Mol. Biol*, **304**, 687-698.

Estrem,S.T. et al. (1998) Identification of an UP element consensus sequence for bacterial promoters. *Proc. Natl. Acad. Sci. U.S.A*, **95**, 9761-9766.

Friedrich,M.J. (2008) Microbiome Project Seeks to Understand Human Body's Microscopic Residents. *JAMA*, **300**, 777-778.

Gama-Castro,S. et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, **36**, D120-124.

Guerrier-Takada,C. et al. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849-857.

Hawkey,P.M. and Jones,A.M. (2009) The changing epidemiology of resistance. *J. Antimicrob. Chemother*, **64 Suppl 1**, i3-10.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563-577.

Humphrey,W. et al. (1996) VMD: visual molecular dynamics. *J Mol Graph*, **14**, 33-38, 27-28.

Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, **3**, 318-356.

Johnson,D.S. et al. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497-1502.

Kruger,K. et al. (1982) Self-splicing RNA: autoexcision and autocyclization of the

ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147-157.

Lai,E.C. (2003) RNA sensors and riboswitches: self-regulating messages. *Curr. Biol*, **13**, R285-291.

Mathew,R. and Chatterji,D. (2006) The evolving story of the omega subunit of bacterial RNA polymerase. *Trends in Microbiology*, **14**, 450-455.

Meng,X. et al. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol*, **23**, 988-994.

Neuwald,A.F. et al. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, **4**, 1618-1632.

Peterson,J. et al. (2009) The NIH Human Microbiome Project. *Genome Res*, **19**, 2317-2323.

Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, **8**, 269-278.

Prakash,A. et al. (2004) Motif discovery in heterogeneous sequence data. *Pac Symp Biocomput*, 348-359.

Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 784 -788.

Schürch,A.C. et al. (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J. Clin. Microbiol*, **48**, 3403-3406.

Serganov,A. (2009) The long and the short of riboswitches. *Curr. Opin. Struct. Biol*, **19**, 251-259.

Siddharthan,R. et al. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates

phylogeny. *PLoS Comput. Biol*, **1**, e67.

Sinha,S. et al. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16-23.

Stormo,G.D. and Hartzell,G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. U.S.A*, **86**, 1183-1187.

Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505-510.

Turnbaugh,P.J. et al. (2007) The human microbiome project. *Nature*, **449**, 804-810.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369-2380.

Zhao,Y. et al. (2009) Inferring Binding Energies from Selected Binding Sites. *PLoS Comput Biol*, **5**, e1000590.

Chapter 2: Sequence-based motif finding protocol in prokaryotic systems[1]

1 This chapter is adapted from **Sahota, G.** & Stormo, G. D. Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. *Bioinformatics* **26**, 2672-7 (2010).

Abstract

Motivation: Computational techniques for microbial genomic sequence analysis are becoming increasingly important. With next-generation sequencing technology and the human microbiome project underway, current sequencing capacity is significantly greater than the speed at which organisms of interest can be studied experimentally. Most related computational work has been focused on sequence assembly, gene annotation, and metabolic network reconstruction. We have developed a method that will primarily use available sequence data in order to determine prokaryotic transcription factor binding specificities.

Results: Specificity determining residues (critical residues) were identified from crystal structures of DNA-protein complexes and transcription factors (TFs) with the same critical residues were grouped into specificity classes. The putative binding regions for each class were defined as the set of promoters for each TF itself (autoregulatory) and the immediately upstream and downstream operons. MEME was used to find putative motifs within each separate class. Tests on the LacI and TetR TF families, using RegulonDB annotated sites, showed the sensitivity of prediction is 86% and 80% respectively.

Availability: http://ural.wustl.edu/~gsahota/HTHmotif/

Introduction

There are more bacterial species than from any other kingdom, but only a few have been studied in much detail. Their relatively small genomes make them readily amenable to sequencing and they now constitute the most abundant genome sequences in

the public databases. Projects such as the Human Microbiome Project (Turnbaugh et al., 2007) and other metagenomic sequencing projects (Riesenfeld et al., 2004) promise to significantly increase the amount of genomic sequence from bacterial species. For most of these species the genome sequence is the only available information so computational approaches are essential to learning more about their characteristics and capabilities. Most current computational analyses focus on sequence assembly (Pop, 2009; Ye and Tang, 2009), the phylogenetic distributions of species (Hamady et al., 2009; Pei et al., 2009), functional classification of gene (Selengut et al., 2010; Qin et al., 2010) and metabolic network reconstruction (Ye and Doak, 2009). Many of these analyses are accomplished through the identification of homologous proteins with known function and the inference of functional conservation in the newly sequenced species. As of yet, there has been little computational work focused on transcriptional regulation in these prokaryotic systems. In this paper, we present a novel sequence-based method to infer the specificities of prokaryotic transcription factors (TFs) through the comparisons of their DNA-binding domains and applying a motif-finding algorithm to likely binding regions.

Most prokaryotic TFs contain an helix-turn-helix (HTH) fold, where the second helix, also known as the recognition helix, primarily contacts DNA (Harrison, 1991; Perez-Rueda and Collado-Vides, 2000; Santos et al., 2009). Using crystal structures of protein-DNA complexes, we can determine a set of residues that is important for defining the specificity of the protein, the “critical residues”. Commonly, these HTH TFs bind as homodimers with palindromic DNA specificities. Previous studies have utilized those features to identify regulatory motifs in related bacterial species but in those cases the TF that binds the motif was not identified except in cases where the motif corresponded to

one for a known TF (McCue et al., 2002; Qin et al., 2003). In general, when the binding motif for a specific TF is known, and orthologous TFs are identified in other species, one can transfer the knowledge about the motif and predict genes that are regulated by the TF in the new species (Alkema et al., 2004; Gelfand et al., 2000b; Tucker et al., 2004; Yu et al., 2004). Making connections between novel motifs and the TFs that bind them can also be accomplished by taking into account additional information (Tan et al., 2005). In that study the most useful information for identifying the TF that bound to a specific motif was the proximity of the TF, within the genome, to the locations of the predicted binding sites. In a similar approach, motifs for orthologous TF were predicted based on the assumption of autoregulation (Sorokin et al., 2009). In an earlier study of the Escherichia coli transcriptome, approximately 55% of the TFs analyzed were estimated to be autoregulated (Martínez-Antonio and Collado-Vides, 2003). Our analysis using RegulonDB 6.7 (Gama-Castro et al., 2008) indicates that this value increases to 78% if one also includes the promoters of neighboring operons.

Motif finders typically depend on having at least one of two types of data. In a "phylogenetic footprinting" approach one has orthologous genes from a set of species and attempts to find the conserved binding site motifs that control their expression (Berezikov et al., 2004; Blanchette and Tompa, 2002; Cliften et al., 2003; Wang and Stormo, 2005). Using such data one can often find potentially functional regulatory motifs, but the TFs that bind to them are frequently unknown. The other general approach uses sets of sequences within one species for which experimental data suggest they contain common binding sites. This may be the promoters (or other regulatory regions) that are known to be regulated by a common TF, or sets of genes that are found to be co-regulated, perhaps

by unknown TFs (Bailey and Elkan, 1994; Buhler and Tompa, 2002; Down and Hubbard, 2005; Hertz and Stormo, 1999; Liu et al., 2001; Pavesi et al., 2001; Thompson et al., 2003). More recently ChIP-chip and ChIP-Seq methods have been used to identify genomic regions that bind to a specific TF. Both kinds of data can be used simultaneously, where sets of genes within one species are thought to be co-regulated, or at least co-bound by the same TF, and one also has the orthologous regions from multiple species from which to focus on the conserved sites (Gelfand et al., 2000a; Jensen et al., 2005; Kellis et al., 2003; Moses et al., 2004; Prakash et al., 2004; Siddharthan et al., 2005; Sinha, 2007; Wang and Stormo, 2003). But for the vast majority of sequenced bacterial species, data that can be used to identify the binding sites for specific TFs is not available. There is generally no experimental data from which to identify co-regulated genes, and frequently bacterial TFs, such as LacI, only regulate one gene so that canonical motif finding would not work. While motifs can be found for orthologous genes across multiple species, that often works only for relatively closely related species and not for the entire distribution of bacterial genomes that are sequenced (Lozada-Chávez et al., 2006; Price et al., 2007). In analyzing metagenomic data, the definition of orthologous genes also becomes quite difficult because one only has partial genome sequences. It also doesn't identify the TF that binds to the motif, which is necessary to be able to begin determining the regulatory networks across bacteria.

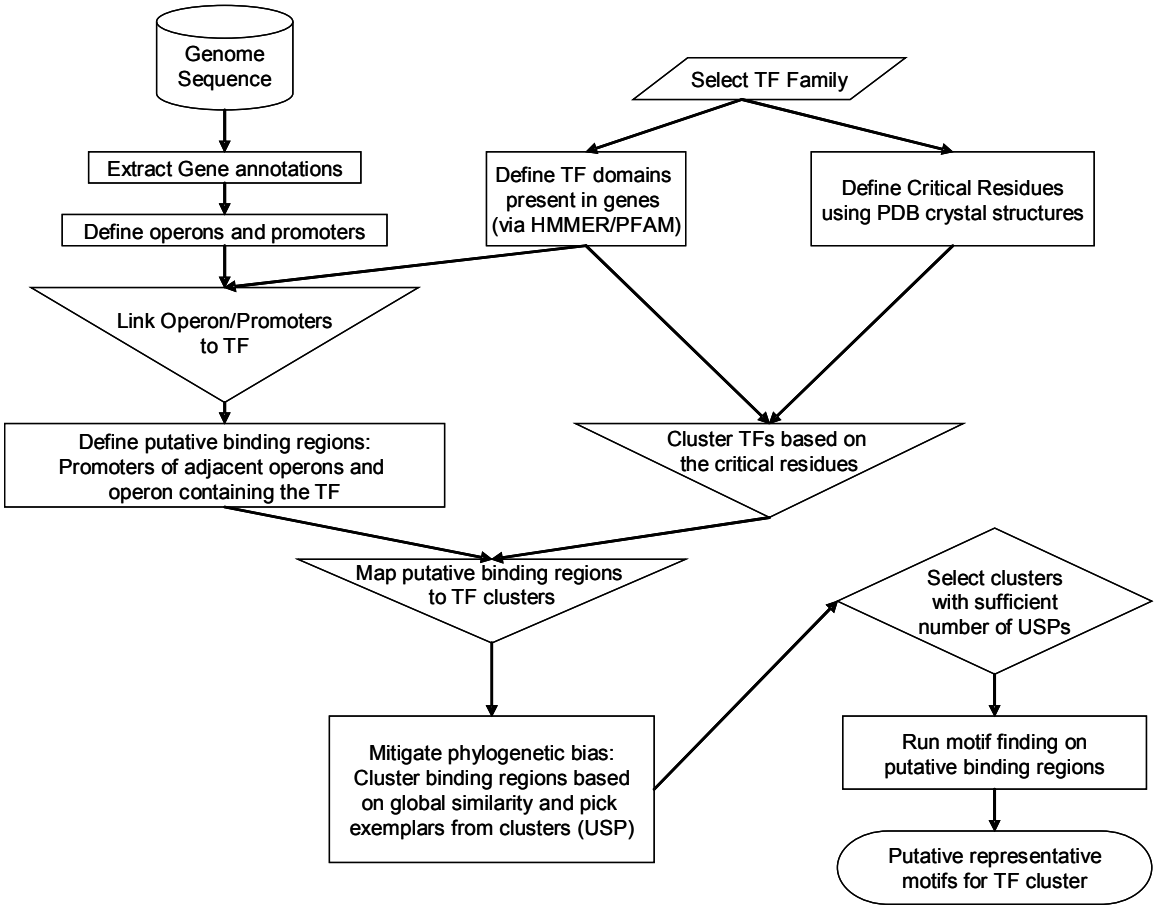
The approach we take in this paper relies on three types of information. The first is the identification of bacterial TFs that contain HTH domains and their classification into subfamilies based on the primary protein sequence signatures within the HTH domain using Pfam (Finn et al., 2010). These subfamilies are not necessarily functionally

related, as many times the functions of a specific protein are determined by a separate effector domain, but proteins within the same subfamily are likely to interact with DNA very similarly, and in particular to use the same critical residues for determining the binding specificity of the TF (Contreras-Moreira and Collado-Vides, 2006; Morozov and Siggia, 2007; Siggers et al., 2005). The second type of information is the structure of the DNA-protein complex for at least one member of the subfamily, which is obtained from PDB (Dutta et al., 2009). There are 22 subfamilies of bacterial HTH TFs that have at least one known crystal structure from which we can determine the protein residues, within the HTH domain, that determine the binding specificity. We cluster together TFs from the same subfamily that also contain the same critical residues as we expect them to bind to identical, or at least very similar, motifs whether or not they are orthologous TFs. The third type of information we need in order to assign motifs to each TF cluster is a set of likely binding regions. For this we rely on the fact, mentioned above, that most bacterial TFs regulate themselves and/or adjacent operons. Therefore we only need short contigs, containing the TF and its promoter as well as adjacent promoters, to have a high likelihood of having regions that contain binding sites for the TFs. Although there are currently many complete bacterial genomes in the database, there are also many genomes represented only by whole genome shotguns (WGS) in which the contigs are considerably shorter, and the large scale microbiome projects that are beginning will also generate large datasets with much smaller contigs. The approach we apply in this paper will be able to leverage that type of data to identify the binding motifs for many uncharacterized bacterial TFs, which opens the way to start modeling regulatory networks. We demonstrate this approach on two HTH subfamilies, TetR and LacI which

are large HTH subclasses with protein-DNA crystal structures and palindromic specificities. Initially, this large size will be important in order to have enough sequences within each specificity class to confidently find motifs.

Methods

Figure 2.1. Flowchart describing overall method. The shapes are standard flowchart shapes, with the disk showing a database, parallelograms showing user input, the rectangle showing processes, inverted triangles showing merging, diamonds showing selection, and the rounded box showing the terminating state.



A conceptual overview of the method is shown in Figure 2.1. The implementation was via a series of Perl scripts, using inline C code for the pairwise sequence alignment

for speed.

Processing genomes

Four main types of genomic data were selected from the NCBI ftp site for this project: completed bacterial genomes, completed plasmid genomes, completed phage genomes and bacterial whole genome shotgun (wgs) projects. For the former three datasets, the genbank DNA sequence (gbk), the protein translation tables (ptt) and rna translation tables (rnt), and the protein fasta sequence (faa) files were retrieved. For the plasmids and the phages (retrieved as viruses), they were filtered to only contain bacterial or phage sequences respectively. In the case of the whole genome shotgun sequences, these subfiles were generated from the genbank flat file (gbff). The downloaded files were validated for correct file format structure and the gbk files were used to recreate the ptt and rnt files when errors were found. Each separate gbk file, with its ptt and rnt annotations, was further processed into operon and promoters using distance cutoffs similar to those used previously (Liu et al., 2008; Price et al., 2005; Tan et al., 2005): a 50 bp intergenic cutoff for determining operon membership and a 400bp maximum promoter length relative the first gene of the operon. The minimum promoter size was required to be 50bp. Circular sequences were appropriately handled both in defining the operons as well as their respective promoters.

Determining critical residues

To classify the proteins into their respective HTH subfamilies, HMMER v3.0rc2 (Eddy, 2009) and Pfam v24.0 were used. The PDB files for each Pfam entry were used to determine the critical residues for that subfamily (ie. residues that contact DNA). This

was achieved using a modified version of PDB2PQR v1.5 (Dolinsky et al., 2007) to find all protein residues within 3.5 angstroms of a DNA base-pair (excluding backbone atoms) that may form van der Waals contacts or hydrogen bonds, where the distance cutoff was 3.4 angstroms and 30 degree maximal angle between the acceptor, donor, and donor hydrogen atoms. A maximum of two hydrogen bonded bridging waters were allowed between the protein and DNA base. The union of all of these sets of potential contacts yielded the critical residues.

Finding TF family members

The HMM for the Pfam entry was used to search through the fasta sequences using hmmsearch. TFs containing multiple domains of an HTH subfamily were removed. The resulting domains were then aligned using hmmalign. In order to try to maintain a similar binding mechanism, no gaps/insertions were allowed in the alignment within the range bounded by the critical residues unless the same gaps/insertions had also been seen in the PDB structures. Critical residues were also constrained to fall within the boundaries of the HMM domain.

Defining putative binding regions

Under the assumption that prokaryotic transcription factors regulate nearby operons, for every TF, the upstream promoter and the two promoters of the neighboring operon in either direction were concatenated to generate a "super-promoter" which potentially contains a binding site for the TF. The WGS sequence data was in the form of contigs rather than finished full length sequences, thus there was no guarantee that all three component promoters would be part of the same contig. In these cases where the

contigs did not contain all of the specified promoters, the subset of present relevant promoters were used to define the super-promoter.

Clustering TFs and generating USPs

For each HTH subfamily, these critical residue sets, which will be referred to as CR tags, were used to cluster the transcription factor protein sequences. These clusters of TFs needed be mapped into clusters of putative binding regions in order to proceed with motif finding. Given the biased distribution of sequenced genomes and the potential for non-unique genomes in the procedure above, a simple mapping of the TFs to their super-promoters would not be sufficient to generate an approporiate dataset for motif finding. Instead, a subset, known as unique super-promoters (USPs), was defined as described.

To compare two super-promoters, the component promoter sequences were compared. To mitigate differences due to promoter shuffling or varying sizes of super-promoters, an all-by-all comparison was performed, using a trimmed Needleman-Wunsch (NW) (Needleman and Wunsch, 1970) alignment (1/-1, -2 score scheme for match/mismatch, gap respectively, excluding the trailing gaps). The alignment score for each pair of promoters was normalized to fall between 0 and 1. The Hungarian algorithm (Kuhn, 2005) optimization was applied to the normalized NW scores to determine the best pairs of related promoter sequences. The “super-promoter” weighting score was defined as the average of the best component pairwise scores. These weighting scores were used to generate a hierarchical complete linkage tree via the perl module Algorithm-Cluster version 1.4.6 and using a threshold of 90% of the theoretical maximum score, this tree was cut to define clusters of sequences. Each resulting cluster of promoter sequences was considered one effective sequence and the sequence closest to the center of the

cluster was chosen as the exemplar. If multiple sequences were equidistant from the center, ties were broken using the length of the sequence and the species of origin. This set of exemplar sequences was the USP set for each TF cluster and used for the motif finding procedure described below.

Motif Finding

MEME v4.3.0 (Bailey and Elkan, 1994) was used to determine the motifs of these USP sets. Only TF clusters with a minimum of 10 USPs were used in motif finding. A maximum of three motifs were reported for each cluster. MEME was run allowing zero or one site per sequence (zoops), the sites were required to be palindromic and the motif width was restricted to be between 15 and 25bp. For further analysis, motifs were required to have an e-value of < 1. The result of motif finding was a set of motifs that likely contained the true representative binding site of the CR tag cluster members.

Validation

All TFs and promoter sequences from Escherichia coli K12 MG1655 (genbank code NC_000913) along with the corresponding promoter sequence cluster members were excluded from the motif discovery sets so that they could be used as independent test sets for evaluating the effectiveness of this procedure to identify true motifs for bacterial TFs. In order to validate the predicted motifs, a Z-score based metric was used to search the sites defined in RegulonDB 6.7. The position weight matrix (PWM) was calculated from the frequency matrix as has been described earlier, using a pseudocount of 1 (Hertz et al., 1990). The weighted mean and variance was determined for each column (position) of the PWM, weighting by the relative background frequencies for

each base. The mean and variance of the PWM were calculated as the sum of the means and variances of the individual columns. The standard deviation was calculated as the square root of the summed variance. Sequences were scored using an additive model as the sum of the position weight matrix elements for each sequence position. The threshold Z-score of 5, corresponding to roughly one match in the E. coli genome by chance, was used to specify whether the RegulonDB site matched the motif. A motif was considered correct if any of the RegulonDB sites for that transcription factor exceeded this threshold. The quality measure is the sensitivity of finding a correct motif, (TP) / (TP+FN).

Results

Between releases 176 and 177, Genbank contained 1056 complete bacterial genomes and 634 whole genome shotgun datasets as well as 2011 plasmid and 543 phage genomes. In this study we have focused on the LacI (PF00356) and TetR (PF00440) Pfam HTH subfamilies (Table 2.1). These two subfamilies comprise roughly 1/10 of the HTH domains in the Pfam clan CL0123. From that set of DNA sequences, for LacI there are 5989 domains. When filtered to remove gaps in restricted positions, multiple domains and missing promoter sequences, there are 5258 TFs remaining, and we defined the critical residues based on three LacI family proteins that have structures bound to DNA: LacI (PDB codes 1efa, 1jwl, 2pe5); CcpA (PDB codes 1rzr, 1zvv); PurR (PDB codes 1bdh,1 bdi, 1jfs, 1jft, 1jh9, 1pnr, 1qp0, 1qp4, 1qp7, 1qpz, 1qqa, 1qqb, 1vpw, 1wet, 1zay, 2pua, 2pub, 2puc, 2pud, 2pue, 2puf, 2pug). Based on those structures, we determine that there are ten critical residues (positions 2,3,12,13,14,17,18,24,25,26) in the respective Pfam HMM domain (PF00356). Using those ten positions to define the specificity classes, there are a total of 1827 classes. There are 23119 TetR domains which, after the

same data filtering as above, lead to 21883 TFs. In TetR, we defined the critical residues based on three proteins that have structures bound to DNA in the PDB: TetR (PDB code 1qpi); QacR (PDB code 1jt0); CGL2612 (PDB code 2yvh). Based on those structures we determine that there are seven critical residues (positions 20,29,30,31,32,34,35) from the respective Pfam HMM domain (PF00440). Using those seven positions to define specificity classes, there are a total of 6207 classes. The distribution of the sequences into these specificity clusters is shown in Figure 2.2.

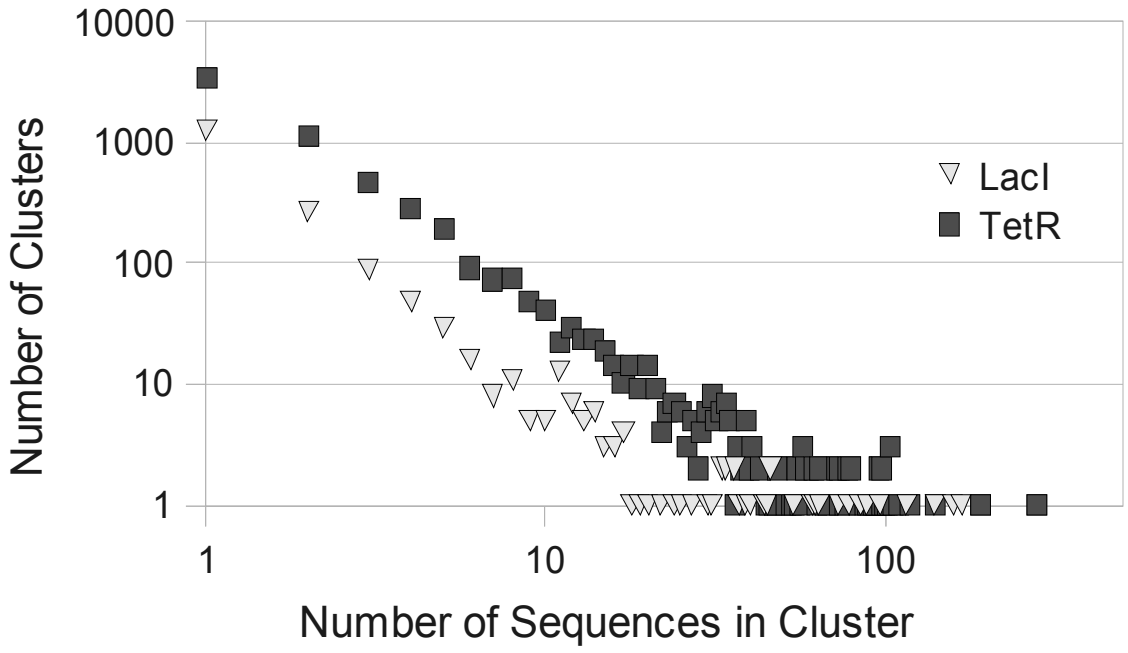
Table 2.1. Dataset sizes for LacI and TetR. The values shown are the number of sequences (with the exception of the first values that are the number of domains), within the parenthesis are the number of specificity clusters. Briefly, USPs are promoter sets that have been filtered to remove redundancy (see methods for more detail). Hamming Distance (HD) is a measure of similarity, the number of substitutions required to change one string into another.

Type	LacI (PF00356)	TetR (PF00440)
Domains	5989	23119
Transcription Factors	5258 (1827)	21883 (6207)
USP >=10 sequences	1733 (32)	8124 (226)
Predicted Motifs	1716 (31)	7923 (214)
Within 1HD of Motif CR tag	1958 (95)	11394 (1335)
Within 2HD of Motif CR tag	2409 (293)	16929 (3801)

The identity of the residues at the critical residue positions define a sequence tag. For example, the cluster TetR-SPKGSYH refers to the set of all proteins that are classified as belonging to the TetR family of HTH proteins and have residues S,P,K,G,S,Y,H in HMM alignment positions 20,29,30,31,32,34,35 respectively. For motif finding, the promoter sequences with potential binding sites were inferred using the computationally derived operons, taking the promoter of the operon containing the TF,

and the upstream and downstream operon promoters as well. These three promoters were concatenated into a "super-promoter" which likely contained at least one binding site for the TF.

Figure 2.2. Log-scale plot showing size distribution of specificity clusters for LacI (inverted triangles) and TetR (squares).



There are a large number of similar genomes that have been sequenced, many times these are simply different strains of model prokaryotic organisms. This introduces the issue of similar sequence due to a similar lineage which could be resolved using a phylogenetic tree. However, there is also the potential issue of horizontal gene transfer as well. In order to resolve the potential promoter redundancy issue, if the promoters were too similar, they were reduced to one exemplar sequence.

In order to ensure stability in motif finding, a minimum of 10 USPs was required for running MEME. With this threshold criteria, there were 32 (1733) and 226 (8124)

classes (TFs) for LacI and TetR, respectively, and putative motifs were obtained for 31 and 214 of them. The majority of classes without motif prediction were simply due to a lack of sufficient USPs to reliably undertake motif finding (Table 2.1).

RegulonDB was used to validate the datasets where E coli was excluded. These datasets comprise a subset of the full dataset and will be described in further detail, however, the predicted motifs and sequence datasets for all of the classes are available at http://ural.wustl.edu/~gsahota/HTHmotif/. The external validation test using RegulonDB showed that in LacI, there were 6 true positives and 1 false negative, for an accuracy of 86% (Table 2.2 and Figure 2.3). For TetR, there were 4 true positives and 1 false negatives, for an accuracy of 80% (Table 2.3 and Figure 2.4). Additionally, we included an analysis of whether the motif was present as part of the superpromoter of the excluded TF to test the hypothesis of local regulation (Tables 2.2 and 2.3). In most instances, it appears to be a valid assumption, but even in TetR-RAPTYSR where this assumption was not true, the protocol was still able to predict the correct motif, due to local regulation in other bacterial organisms in the same cluster.

Table 2.2. Validation results for LacI using RegulonDB 6.7. The matches in RegulonDB are coded as follows: Y means a match to any site, N means no matches and – means no RegulonDB site. The autoreg column is a + if there is a match to the superpromoter and - if there is no match.

Locus ID	Name	CR Tag	Sequences	USP	Matches in RegulonDB	Autoreg
b1658	purR	IKSTTSHRFV	168	60	Y	+
b2837	galR	IKSVASRPKA	140	45	Y	+
b3753	rbsR	MKSTSSHRFV	116	41	Y	+
b4241	treR	IKGKSSRSGV	97	15	Y	+
b0345	lacI	LYSYQSRSHV	37	14	Y	+
b2714	ascG	MLSKASRGYV	62	11	Y	+
b3934	cytR	MKSTASRDKV	85	12	N	+
b1320	ycjW	IYSKSSRTNI	61	20	-	+

Table 2.3. Validation results for TetR using RegulonDB 6.7. The matches in RegulonDB are coded as follows: Y means a match to any site, N means no matches and – means no RegulonDB site. The autoreg column is a + if there is a match to the superpromoter and - if there is no match.

Locus ID	Name	CR Tag	Sequences	USP	Matches in RegulonDB	Autoreg
b3963	fabR	RAPTSYR	193	81	Y	-
b1649	nemR	SPKGSYH	141	49	Y	+
b0313	betI	ASTGISH	98	41	Y	+
b3264	envR	NTRGAYW	104	27	Y	+
b1013	rutR	ESKTNLY	95	18	N	+
b3641	slmA	ASEAAYR	282	148	-	+
b0846	ybjK	RPLGSTY	118	45	-	+
b4251	yjgJ	ANPPSYA	87	19	-	+
b0796	ybiH	RNIATTY	105	17	-	+
b1111	ycfQ	AKAPTYA	97	17	-	+

Discussion

Using primarily genomic sequence data augmented with structural priors, we are able to determine putative motifs for a number of bacterial TFs in two families. The method described is capable of working not only with fully sequenced genomes, but also with sufficiently long contigs, allowing for the use of assembled metagenomic reads. Using the MEME program, putative motifs were determined for 31 (LacI) and 214 (TetR) classes of TFs representing ~1/3 of the sequences of each of these TF families. For validation, classes that had an excluded E coli USP were selected, 8 (LacI) and 10 (TetR). Of these selected classes, 7 (LacI) and 5 (TetR) had known regulatory sites in RegulonDB. The majority of these motifs, 6/7 for LacI and 4/5 for TetR, were consistent with the known regulatory sites defined based on RegulonDB, validating this approach. Even some of the motifs that did not match to known sites with scores exceeding our stringent threshold still had fairly high scores and are likely to be very similar to the true

motifs for those classes.

While this approach has proven to be useful, there are several modifications to the method we describe in this paper that should offer further improvements in our ability to determine binding motifs for bacterial TFs. The current protocol assumes a fixed-width gap between the half-sites of the motif. However, there is no guarantee that proteins with similar critical residues must have similar gaps in the spacer region between the half-sites, as the regions of the protein that determine these variable gaps are generally outside of the DNA binding domain (Laguri et al., 2003; Mao et al., 2005; Reece and Ptashne, 1993). Even within the test set, we can see evidence of multiple widths in TetR-SPKGSYH. In the first and third motif, there is a TAGACC half site, separated by a 4 or 0 base spacer from the complementary GGTCTA. For TetR-NPKGSYH, these spacers are 3 or 0 bases. An EM-based algorithm that allowed variable spacing between the two parts of E. coli promoters had been published previously (Cardon and G D Stormo, 1992) and a similar approach could increase the power of detecting some motifs that may be missed simply because they have multiple binding widths. Current gapped motif finders are not capable of dealing with sequence fragments, which is important in the context of the “super-promoters”.

We only applied MEME to classes with at least ten USPs because motif finding is more reliable with larger datasets. However, many of the classes with less than ten USPs are very similar to other classes with ten or more, and we expect that TFs with very similar critical residues will bind to very similar motifs. This means that we could use the motifs from the larger classes as priors to aid in the discovery of the motifs for the smaller classes. As shown in Table 2.1, there are an additional 95 and 1335 classes that

are at a Hamming distance of one (HD=1) away from the larger classes for the LacI and TetR families, respectively. If we go to HD=2 the number of classes increases to 293 (LacI) and 3801 (TetR). This could greatly increase the number of TF classes for which motifs could be determined and further expand the repertoire of TF-motif pairs. In the current implementation, a set of putative motifs is predicted for the TF cluster, however the correct motif within the set is not specified. In conjunction with the above described gapped motif finder, these HD classes could also be used to filter out inconsistent or incorrect motif predictions, under the assumption that similar CR tags lead to similar half-sites or motifs. This refined analysis would lead to a one-to-one mapping of a predicted motif to a TF cluster. Another benefit of having a large number of TF-motif pairs is the determination of the interacting residues. Our choice of the critical residues is based on crystal structures of DNA-protein complexes where we have used a distance cutoff between an amino acid and a base-pair to identify those residues that may, in at least some members of the family, contribute to the specificity of binding. It has been shown before that interface residues are only partially conserved across DNA binding domains (Contreras-Moreira et al., 2010). In this paper, the union of all such residues was used, leading to a potential overspecification of the critical residue set, in turn decreasing the size of certain classes. In addition, it may be that some residues, while close to the DNA, do not participate in binding specificity and could be eliminated from the critical residue set, which would increase the size of the classes. In general, correlations between the aligned protein sequences and alignments of the motifs can be useful in determining which protein residues interact with which base-pairs (Mahony et al., 2007; Noyes et al., 2008). This could then be used to determine the critical residues even for TF families

currently without crystal structures for DNA-protein complexes.

Finally, there are many more HTH families that can be addressed with this approach. HTH proteins are classified as the clan CL0123 in Pfam and there are 141 HTH subfamilies of which 22 are mainly bacterial domains and contain protein-DNA crystal structures where the domain interacts with DNA. When taking into account the variability in size of the families, this actually covers approximately ½ of the potential HTH TF proteins, so the method has significant potential to cover a large amount of the potential HTH TF proteins. In some of these families, we may have motifs with variable gaps in their spacer regions and differing configurations of the half-sites such as direct repeats instead of palindromic motifs, and sometimes even mixtures of the two modes. This will require a modification to the current protocol but may provide for a much larger collection of binding motifs for specific bacterial TFs.

Acknowledgements

We thank all members of the Stormo lab for helpful discussions and advice about this work.

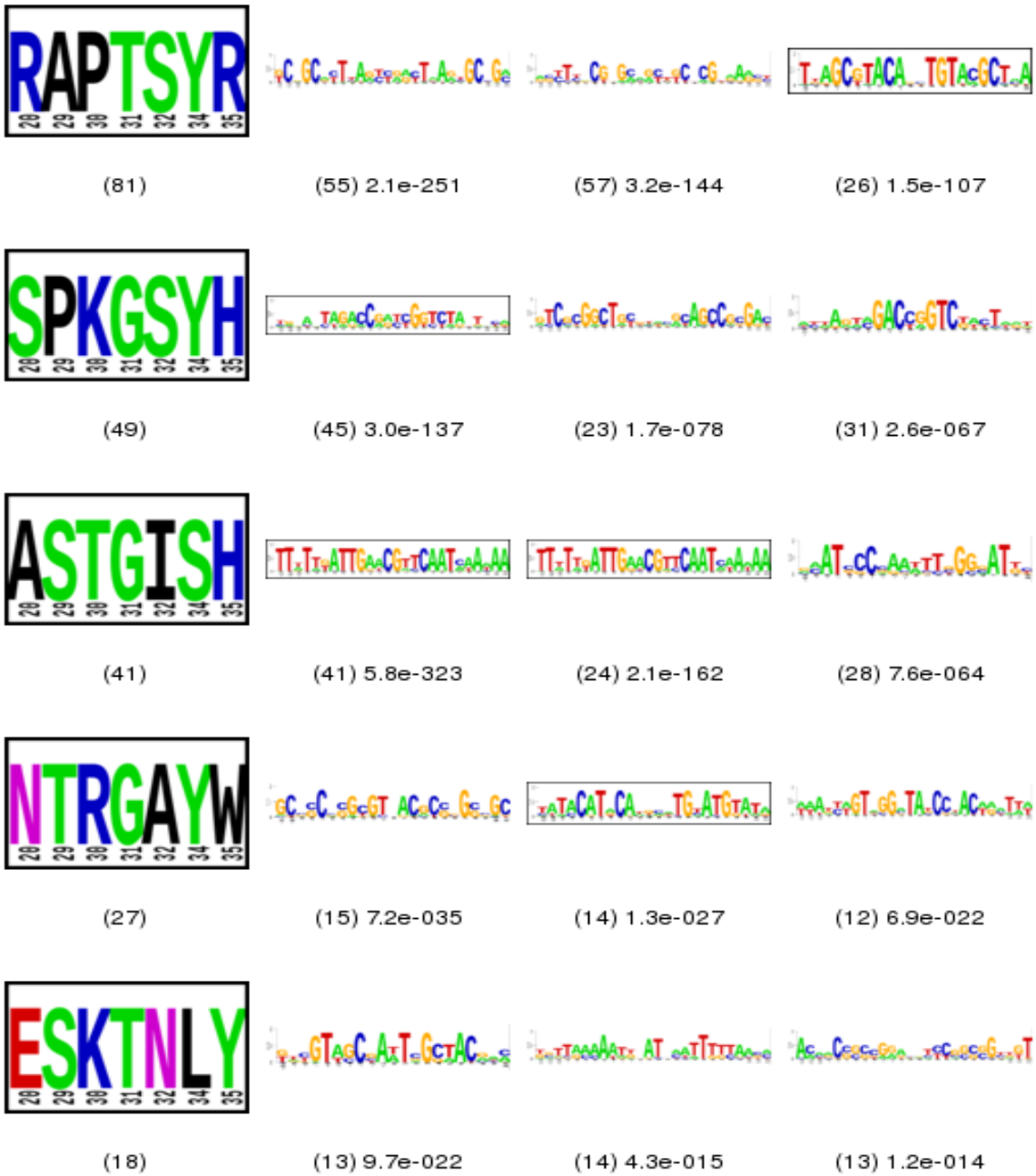
Funding: National Institutes of Health [T32 GM07200, T32 GM008802 to G.S., and R01 HG00249 to G.D.S].

Conflict of Interest: None declared

Supplementary Figures

Figure 2.3. Visual representation of the validation results for LacI. RegulonDB matches are shown by black boxes, surrounding both the CR tag and the matching motifs. In parenthesis are the number of USPs, below the CR tag or the number of sites, below the motifs. Additionally below the motifs, the MEME e-value is shown.

Figure 2.4. Visual representation of the validation results for TetR. RegulonDB matches are shown by black boxes, surrounding both the CR tag and the matching motifs. In parenthesis are the number of USPs, below the CR tag or the number of sites, below the motifs. Additionally below the motifs, the MEME e-value is shown.



References

Alkema,W.B.L. et al. (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. Genome Res, 14, 1362-1373.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol, 2, 28-36.

Berezikov,E. et al. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res, 14, 170-178.

Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res, 12, 739-748.

Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. J. Comput. Biol, 9, 225-242.

Cardon,L.R. and Stormo,G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. J. Mol. Biol, 223, 159-170.

Cliften,P. et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science, 301, 71-76.

Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. Bioinformatics, 22, e74-80.

Contreras-Moreira,B. et al. (2010) Comparison of DNA binding across protein superfamilies. Proteins, 78, 52-62.

Dolinsky,T.J. et al. (2007) PDB2PQR: expanding and upgrading automated preparation

of biomolecular structures for molecular simulations. Nucl. Acids Res., 35, W522-525.

Down,T.A. and Hubbard,T.J.P. (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res, 33, 1445-1453.

Dutta,S. et al. (2009) Data deposition and annotation at the worldwide protein data bank. Mol. Biotechnol, 42, 1-13.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform, 23, 205-211.

Finn,R.D. et al. (2010) The Pfam protein families database. Nucleic Acids Res, 38, D211-222.

Gama-Castro,S. et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res, 36, D120-124.

Gelfand,M.S. et al. (2000a) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucleic Acids Res, 28, 695-705.

Gelfand,M.S. et al. (2000b) Comparative analysis of regulatory patterns in bacterial genomes. Brief. Bioinformatics, 1, 357-371.

Hamady,M. et al. (2009) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J, 4, 17-27.

Harrison,S.C. (1991) A structural taxonomy of DNA-binding domains. Nature, 353, 715-719.

Hertz,G.Z. et al. (1990) Identification of consensus patterns in unaligned DNA sequences

known to be functionally related. Comput. Appl. Biosci., 6, 81-92.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15, 563-577.

Jensen,S.T. et al. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. Bioinformatics, 21, 3832-3839.

Kellis,M. et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423, 241-254.

Kuhn,H.W. (2005) The Hungarian method for the assignment problem. Naval Research Logistics, 52, 7-21.

Laguri,C. et al. (2003) Solution structure and DNA binding of the effector domain from the global regulator PrrA (RegA) from Rhodobacter sphaeroides: insights into DNA binding specificity. Nucl. Acids Res., 31, 6778-6787.

Liu,J. et al. (2008) The cis-regulatory map of Shewanella genomes. Nucleic Acids Res, 36, 5376-5390.

Liu,X. et al. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138.

Lozada-Chávez,I. et al. (2006) Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res, 34, 3434-3445.

Mahony,S. et al. (2007) Inferring protein DNA dependencies using motif alignments and mutual information. Bioinformatics, 23, i297-304.

Mao,L. et al. (2005) Combining microarray and genomic data to predict DNA binding motifs. Microbiology, 151, 3197-3213.

Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol, 6, 482-489.

McCue,L.A. et al. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res, 12, 1523-1532.

Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. Proc. Natl. Acad. Sci. U.S.A, 104, 7068-7073.

Moses,A.M. et al. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Pac Symp Biocomput, 324-335.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48, 443-453.

Noyes,M.B. et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell, 133, 1277-1289.

Pavesi,G. et al. (2001) An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics, 17 Suppl 1, S207-214.

Pei,A. et al. (2009) Diversity of 23S rRNA genes within individual prokaryotic genomes. PLoS ONE, 4, e5437.

Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. Nucl. Acids Res., 28, 1838-1847.

Pop,M. (2009) Genome assembly reborn: recent computational challenges. Brief Bioinform, 10, 354-366.

Prakash,A. et al. (2004) Motif discovery in heterogeneous sequence data. Pac Symp

Biocomput, 348-359.

Price,M.N. et al. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucl. Acids Res., 33, 880-892.

Price,M.N. et al. (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput. Biol, 3, 1739-1750.

Qin,J. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, 59-65.

Qin,Z.S. et al. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. Nat Biotech, 21, 435-439.

Reece,R.J. and Ptashne,M. (1993) Determinants of binding-site specificity among yeast C6 zinc cluster proteins. Science, 261, 909-911.

Riesenfeld,C.S. et al. (2004) METAGENOMICS: Genomic Analysis of Microbial Communities. Annu. Rev. Genet., 38, 525-552.

Santos,C.L. et al. (2009) A phylogenomic analysis of bacterial helix-turn-helix transcription factors. FEMS Microbiology Reviews, 33, 411-429.

Selengut,J. et al. (2010) Sites Inferred by Metabolic Background Assertion Labeling (SIMBAL): adapting the Partial Phylogenetic Profiling algorithm to scan sequences for signatures that predict protein function. BMC Bioinformatics, 11, 52.

Siddharthan,R. et al. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput. Biol, 1, e67.

Siggers,T.W. et al. (2005) Structural alignment of protein--DNA interfaces: insights into the determinants of binding specificity. J. Mol. Biol, 345, 1027-1045.

Sinha,S. (2007) PhyME: a software tool for finding motifs in sets of orthologous sequences. Methods Mol. Biol, 395, 309-318.

Sorokin,V. et al. (2009) Systematic prediction of control proteins and their DNA binding sites. Nucleic Acids Res, 37, 441-451.

Tan,K. et al. (2005) Making connections between novel transcription factors and their DNA motifs. Genome Res, 15, 312-320.

Thompson,W. et al. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res, 31, 3580-3585.

Tucker,N.P. et al. (2004) DNA binding activity of the Escherichia coli nitric oxide sensor NorR suggests a conserved target sequence in diverse proteobacteria. J. Bacteriol, 186, 6656-6660.

Turnbaugh,P.J. et al. (2007) The Human Microbiome Project. Nature, 449, 804-810.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics, 19, 2369-2380.

Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc. Natl. Acad. Sci. U.S.A, 102, 17400-17405.

Ye,Y. and Doak,T.G. (2009) A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. PLoS Comput Biol, 5, e1000465.

Ye,Y. and Tang,H. (2009) An ORFome assembly approach to metagenomics sequences analysis. J Bioinform Comput Biol, 7, 455-471.

Yu,H. et al. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res, 14, 1107-1118.

Chapter 3: Implementation of a gapped motif finder

Introduction

Computational motif finding involves optimizing a motif model with respect to the data, normally a set of sequences that likely contain DNA binding sites of interest. Many of these computational approaches make differing assumptions regarding the internal structure and distribution of the binding sites within the sequences. Generally, these assumptions are converted into the form of a likelihood function that is optimized. The congruence of these assumptions and the actual parameters underlying the data in part determines the success of the computational motif finders at detecting the true motif (Tompa et al., 2005). Given the specific nature of analyzing TFBS in the context of multiple prokaryotic genomes, a novel motif finder needs to be developed that can handle the motif structures and distributions that are relevant.

The structure of prokaryotic motifs generally consist of palindromic motifs comprised of two half-sites, due to the dimeric quaternary structure of the component transcription factors. It has also been shown before that prokaryotic transcription factors can have a variable gap between these two half-sites (Eraso and Kaplan, 2009). Most motif finders generally have fixed width models and are not built to handle variable gap spacing between two motif half-sites, however there has been some previous work in this area (Bi and Rogan, 2004; Cardon and Stormo, 1992; A. V. Favorov et al., 2005; GuhaThakurta and Stormo, 2001; X. Liu et al., 2001). This problem is exacerbated when one clusters multiple TFs together that may each have slightly different spacing preferences. These spacing preferences can be due to domains or proteins that are outside of the TF binding domain and thus may not be easy to determine *a priori*. Additionally, in most motif finders, the relative half-site orientation is fixed, whereas in bacterial

systems, sometimes the transcription factors bind direct repeats and other times they bind inverted repeats (Bi et al., 2008). A bipartite model of motif finding should allow for a direct repeat (dr), inverted repeat (ir), everted repeat (er) and reverse direct repeat (rdr) as shown in Figure 3.1.

Figure 3.1. Bipartite structure. Shows four relative half-site orientations. Inverted repeats are also commonly known as palindromic motifs.

The distribution model of motifs is generally uniform and continuous across a sequence. In our problem, the more important variable is the continuity of the sequence. There are several possible sequence fragments within which we require expect zero or more motifs. The easiest method to conceptualize this is to consider them as a single fragmented sequence, which will simply be referred to as a sequence in the rest of this chapter. This could be simply implemented as a series of null characters (N, X, -) between each fragment, however, this method would slightly skew the normalization parameters based on sequence length. A more accurate method of implementing sequence fragments is to track the start and end points of the sequence fragments and appropriately handle the edge-case for each fragment similar to how the edge-case would be handled for a single sequence. The difference is that the summation/multiplication of the likelihood function would be over the set of fragments that likely contain a motif, or a

sequence rather than a single unfragmented sequence that likely contains a motif.

Even with an appropriate model, it is necessary to optimize the parameters to maximally fit the data. There are many different methods that can be used in order to perform this optimization. As discussed earlier, there are three main methods that are used in motif finding, greedy algorithms, expectation-maximization (EM) and gibbs sampling. The advantages of greedy algorithms is that they are very fast to run, however, the solution is highly context dependent; the order in which the data is presented to a greedy algorithm can impact the final result. EM algorithms do not suffer this context dependence, however they are only a local optimizer, thus they can be trapped in local energy wells rather than finding the global optimum. Gibbs samplers attempt to solve this problem local energy well by randomly selecting the next starting point using a roulette wheel selection of the posterior probability distribution. In this problem, a deterministic algorithm is preferable and given a reasonable starting model, EM is the best option for a motif optimizer. Additionally, building the motif finding pipeline with an option to optimize a starting motif allows for easier incorporation of prior knowledge of the motif composition. In the context of this dataset, there are CR tags that contain few USPs to reliably attempt *de novo* motif finding, but optimization from a known similar motif may be an alternative option for elucidating the binding motif in these cases.

Before running EM, the parameter values need to be initialized. Given the local optimization nature of the algorithm, the starting parameter values will determine which parameter set, corresponding to a local energy minima, is returned. EM can be implemented with multiple random restarts in order to try to overcome this potential

hurdle, however this solution does not guarantee that the global energy minima will be attained. Alternatively, a set of representative starting points can be used in order to improve the likelihood of finding the global minima.

In the context of motif finding, a set of start points can be defined from the data, generally by some sort of word counting and simple optimization procedure. From this set of start points, a representative subset or potentially all can be used for motif finding. The end result is a set of putative motifs from which the most likely motif must be selected. Selecting between the sets of putative motifs requires a metric to determine which motif is more significant in the context of the data.

The remaining parts of this chapter will discuss the development, preliminary testing and results of a EM-based motif finding method. This motif finder is built on a motif half-site model, where these half-sites are allowed to be in any bipartite orientation and are separated by a variable gap length. In addition, this motif finder is inherently capable of dealing with sequence fragments in a mathematically correct framework, rather than some of the crude approximations described above. The results are preliminary and show that this motif finder is capable of converging to the correct solution given a close enough starting point, which could potentially be a motif from a neighboring cluster that differs slightly in the transcription factor protein specificity tag.

Methods

Overview

The motif finding algorithm consists of two main steps that have been kept separated for reasons of efficiency and modularity. The first step is taking the dataset and

defining a reasonable start point, basically a preliminary motif that can be refined. The second step of the algorithm is to refine this preliminary motif using an EM based algorithm. The input to this second step can include priors on the gap size, the number of expected motifs, the expected relative orientations of the half-sites and obviously the preliminary motif to be optimized. The result of the second step is a predicted motif for the dataset along with a probability distribution of the possible gap sizes.

Start points

Given the additional variables of orientation and gap, a simple enumerative motif finder is run to determine a single optimal start point, which eliminates the requirement of relative motif comparison in a later step. For this algorithm, two parameters are required, the motif width l, and the number of variable positions in the motif d. This is similar to the planted motif problem (Pevzner and Sze, 2000), but in this case for each putative motif, the d different positions are chosen *a priori* rather than having a hamming distance criteria of d potential changes over the entire motif from a fixed consensus.

Given the short width of the half-sites of interest generally on the order of 6-8 bp, instead of generating a set of potential lmers from the input sequences and then clustering them into putative motifs, an enumerative approach was employed. The enumerated set of motifs were of width l containing d variable positions. This process can be broken into two main stages, selecting the d variable positions and generating all possible DNA strings for the (l-d) fixed positions. For implementation efficiency, this was done in the reverse order, as the (l-d) mers would be constant for every (l choose d) combination. After generating all possible (l-d) mers, the d varying positions were chosen and N's were inserted at those d positions. When being added to the final motif set, an orientation

independence criteria was imposed, such that a motif and its reverse complement would not both be part of this finalized set. For the purposes of this chapter, the l,d was fixed as 7,2.

This final motif search set was used to search the input sequences. The bipartite search model including the variable gap was used to search with these enumerated motifs. In the current implementation, the motifs were converted into regular expressions, where Ns were allowed to be A, C, G or T, and these expressions were used to search the sequence fragments. Ranking was based on the number of sequences in which a dyad was found, the number of dyads, the number of sequences in which a monad was found, and the number of monads. These counts were adjusted such that if a motif had overlapping repeats (for example, a polyA tract), the set of identical matches that overlapped by at least one base were only counted a single time and likewise for the dyad. The motif that ranked the highest was chosen as the putative motif. This regular expression motif was transformed into a position frequency matrix (PFM) where the consensus base was given a weight of 0.777 and the other letters are weighted at 0.077 and at the variable positions all nucleotides are given equiprobably weights of 0.25. This resulting PFM was normalized to guarantee that each column summed to 1.

Start points for similar motif optimization

Given a known similar motif, sometimes it is necessary to use this information to guide motif finding in a similar TF cluster, where similarity is judged by the hamming distance between the CR tags. In this case, one could simply take the resulting PWM of the prior motif finding procedure and use that as a start point, however, this runs into the problem that sometimes the predicted motifs have completely conserved bases. These

completely conserved bases are a very high energy barrier for this EM algorithm to overcome, and thus need to be attenuated. A logical method to overcome this problem is to attenuate the entire PWM, by simply adding a pseudocount of 0.1 to every base and then normalizing the resulting PWM to enforce a column sum of 1. This would effectively reset a base that was fully conserved at a value of 1 and 0 for all other bases to roughly the 0.777/0.077 consensus/nonconsensus split as described in the section above.

EM algorithm with variable spacing and multiple binding modes

The basic concept behind the EM algorithm with variable spacing was built on previous work (Cardon and Stormo, 1992). There are four major extensions or modifications to that prior work: calculating probabilities in log-space, allow for a zero-or-one per sequence (ZOOPs) model, correctly working with sequence fragments and bipartite half-site orientations (including palindromic modes). The final program was re-implemented in C to alleviate some restrictions due to the design and architecture of the prior version. Additionally, there is a beta version that was used to prototype some of the advanced features that was written in PERL, however it does not contain the log-space calculations. The results discussed in this chapter have been obtained using the C version.

Before delving into the specific equations and parameters of the method, a brief overview is presented. The probability of a sequence given a motif is shown in Equation 3.1. This probability basically states that the probability of a sequence is composed of bases that are part of the motif, the P(motif) and the bases that are not P(bkgd). In the specific instance of a variably gapped motif, the gap bases are not well defined as being

part of the motif or part of the background. In the prior implementation by Cardon, these were treated separately, however in this implementation, they are simply assumed to be part of the background distribution. Given the probable application of the method, the number of parameters that needed to be estimated from the data would have been increased significantly if the gaps were considered separately. In addition, the relative correspondence of the gap positions in different orientations of binding would have been problematic. If there is no motif present, then the probability is simply the cumulative background probability distribution for all of the bases in that sequence as shown in equation 3.2.

$$P(sequence|motif) = P(motif)P(bkgd_{nonmotif}) \tag{3.1}$$

$$P(sequence|no\_motif) = P(bkgd_{cum}) \tag{3.2}$$

Table 3.1. Variables and definitions.

n	Sequence number
N	Number of sequences
n_f	Number of fragments in sequence n
b b'	base reverse complement of the base
I()	Indicator function that quantity in parenthesis is true (1 or 0)
$I_{n,j,b}$	Indicator function base b at sequence n at position j (1 or 0)
L_n	Sequence length
K	Set of valid start points
Δ	Binding site
Ø	The background
$Y_{n,k}$	Indicator of a Δ starting at position k in sequence n (1 or 0)
m m_1, m_2	Binding orientation (composed of m_1, m_2), hs1 orientation, hs2 orientation (defined as: 0 = forward, 1=reverse)
M	Set of possible binding orientations (maximum of 4: dr, ir, er, idr)

g	Gap size
G	Set of possible gap sizes
$\rho_{b,j}$	Half-site PWM
$\rho_{b,\emptyset}$	Background distribution
ρ_g	Gap size probability
Q	Random variable for binding site presence/absence (1/0)
γ	probability that a sequence contains a binding site
w_{hs}	Half-site width
w	Motif width (w_{hs} + g + w_{hs})

The following sets of equations assume that K is set such that it does not violate the inter-fragment or end-sequence boundaries, and that k is drawn from this set K. Given a specific sequence n and gap size g, the number of valid start positions is shown in equation 3.3.

$$|K| = L_n - (n_f * w) + n_f \tag{3.3}$$

Equation 3.4 shows the probability of a binding site within a specific sequence. The equation can be split into four components, the two half-sites, and then split whether the half-site is in the forward orientation or reverse. Since it is only going to be in one of those two orientations, and given the definition of the orientation as 1 or 0, only one of the two products in the square brackets will be non-zero. And equation 3.5 shows the consequent background distribution assuming a motif that starts at k.

$$\begin{aligned} P(S_{n,\Delta}|Y_{n,k}=1,G=g,M=m) &= [(1-m_1)\prod_{j=k}^{k+w_{hs}}\rho_{b,j}^{I_{n,j,b}} + (m_1)\prod_{j=k}^{k+w_{hs}}\rho_{b,j}^{I_{n,(w_{hs}-j),b'}}] * \\ & \quad [(1-m_2)\prod_{j=k+w_{hs}+g}^{k+w}\rho_{b,j}^{I_{n,j,b}} + (m_2)\prod_{j=k+w_{hs}+g}^{k+w}\rho_{b,j}^{I_{n,(w_{hs}-j),b'}}] \end{aligned} \tag{3.4}$$

$$P(S_{n,\oslash}|Y_{n,k}=1,G=g,M=m) = \prod_{j=1}^{k}\rho_{b,\oslash}^{I_{n,j,b}}\prod_{j=k+w_{hs}}^{k+w_{hs}+g}\rho_{b,\oslash}^{I_{n,j,b}}\prod_{j=k+w}^{L_n}\rho_{b,\oslash}^{I_{n,j,b}} \tag{3.5}$$

The probability of the sequence, given that it contains a motif, is the product of the background and the motif as outlined in equation 3.1 and formally in equation 3.6. Expanded out into its full form in equation 3.7, it becomes clear how the background and motif are distributed across the each position. In addition, the use of the background model for the gap positions also becomes apparent.

$$P(S_n|Y_{n,k}=1,G=g,M=m) \quad = \quad P(S_{n,\triangle}|Y_{n,k}=1,G=g,M=m) * P(S_{n,\oslash}|Y_{n,k}=1,G=g,M=m) \tag{3.6}$$

$$\begin{aligned} P(S_n|Y_{n,k}=1,G=g,M=m) \quad = \quad & \prod_{j=1}^{k} \rho_{b,\oslash}^{I_{n,j,b}} * \\ & [(1-m_1) \prod_{j=k}^{k+w_{hs}} \rho_{b,j}^{I_{n,j,b}} + (m_1) \prod_{j=k}^{k+w_{hs}} \rho_{b,j}^{I_{n,(w_{hs}-j),b'}}] * \\ & \prod_{j=k+w_{hs}}^{k+w_{hs}+g} \rho_{b,\oslash}^{I_{n,j,b}} * \\ & [(1-m_2) \prod_{j=k+w_{hs}+g}^{k+w} \rho_{b,j}^{I_{n,j,b}} + (m_2) \prod_{j=k+w_{hs}+g}^{k+w} \rho_{b,j}^{I_{n,(w_{hs}-j),b'}}] * \\ & \prod_{j=k+w}^{L_n} \rho_{b,\oslash}^{I_{n,j,b}} \end{aligned} \tag{3.7}$$

Since the current model assumes a ZOOPs distribution, there needs to be a null model if there are zero motifs present in the sequence as alluded to by equation 3.2. That is accomplished by assuming that all sequence positions are drawn from the background, shown in equation 3.8.

$$P(S_n|Y_{n,k}=0,G=g,M=m) \quad = \quad \prod_{j=1}^{L_n} \rho_{b,\oslash}^{I_{n,j,b}} \tag{3.8}$$

The discussion has been surrounding the probability of the sequence, however, in the context of the motif finder, the important question is what is the probability that the

binding site starts at position k. Using Bayes theorem on conditional probability shown in equation 3.7, the resulting inversion of the conditional probability is shown is shown in equation 3.9. The P(Q=1) represents the probability that the sequence contains a motif which is parameterized as γ. The Z is the partition function that describes the total probability density and is shown in equation 3.10. The probability in equation 3.9 is bounded between 1 and 0, where 1 means that the binding site is perfectly predicted at only position k and consequently it will be 0 for all other k. Given the ZOOPs model, the sum for all k in equation 3.9 is also bounded between 1 and 0.

$$P(Y_{n,k}=1|S_n,G=g,M=m) = \begin{array}{l} \frac{P(Q=1)*P(S_n|Y_{n,k}=1,G=g,M=m)*P(Y_{n,k}=1|G=g,M=m)*P(G=g|M=m)*P(M=m)}{Z} \\ \frac{P(Q=1)*P(S_n|Y_{n,k}=1,G=g,M=m)*P(Y_{n,k}=1)*P(G=g)*P(M=m)}{Z} \\ \frac{\gamma*P(S_n|Y_{n,k}=1,G=g,M=m)*\frac{1}{|K|}*\rho_g*\frac{1}{|M|}}{Z} \end{array} \tag{3.9}$$

$$\begin{aligned} Z &= P(Q=0)*P(S_n|Y_{n,k}=0,G=g,M=m)+ \\ &\quad \sum_{k\in K} P(Q=1)*P(S_n|Y_{n,k}=1,G=g,M=m)*\frac{1}{|K|}*\rho_g*\frac{1}{|M|} \\ &= (1-\gamma)*\prod_{j=1}^{L_n}\rho_{b,\oslash}^{I_{n,j,b}}+\sum_{k\in K}\gamma*P(S_n|Y_{n,k}=1,G=g,M=m)*\frac{1}{|K|}*\rho_g*\frac{1}{|M|} \end{aligned} \tag{3.10}$$

Equations 3.9 and 3.10 form the basis of the expectation step (E-step) of the EM algorithm. In the maximization step (M-step), the maximum likelihood estimate is calculated as shown in equations 3.11-3.14. The additional superscript of i denotes the value is updated for the i-th iteration.

$$\rho_{b,j}^{(i)} = \frac{\sum_{m\in M}\sum_{g\in G}\sum_{n=1}^{N}\sum_{k\in K} I_{n,(j+k-1),b}*I(S_{n,(j+k-1)}\in\triangle))P(Y_{n,k}=1|S_n,G=g,M=m)}{N} \tag{3.11}$$

$$\rho_{b,\oslash}^{(i)} = \frac{\sum_{m\in M}\sum_{g\in G}\sum_{n=1}^{N}\frac{\sum_{k\in K}\sum_{jin\oslash_k} I_{n,j,b}P(Y_{n,k}=1|S_n,G=g,M=m)}{|K|}}{N} \tag{3.12}$$

$$\rho_g^{(i)} = \frac{\sum_{m\in M}\sum_{n=1}^{N}\sum_{k\in K}P(Y_{n,k}=1|S_n,G=g,M=m)}{\sum_{g\in G}\sum_{m\in M}\sum_{n=1}^{N}\sum_{k\in K}P(Y_{n,k}=1|S_n,G=g,M=m)} \tag{3.13}$$

$$\gamma^{(i)} = \frac{\sum_{g \in G} \sum_{m \in M} \sum_{n=1}^{N} \sum_{k \in K} P(Y_{n,k} = 1 | S_n, G = g, M = m)}{N} \quad (3.14)$$

These equations neglect the priors that can be placed on both ρ_g and γ. The method of implementing the weighted prior is by a normalized proportional weighting of the prior and the updated value and setting the updated value to this sum. Additionally, all of the calculations in the E-step are performed in log space including the summation of logs where necessary.

The EM algorithm is iterated until one of two conditions is met. Either, the maximum number of iterations is reached or the change between the successive iterations, as measured by the absolute difference between the matrices, does not change significantly (10e-16 in the current implementation). Additionally, there is a pseudocount that is applied to the $\rho_{b,j}$ matrix until either a certain number of iterations are performed (600) or until the EM algorithm has sufficiently converged (10e-10).

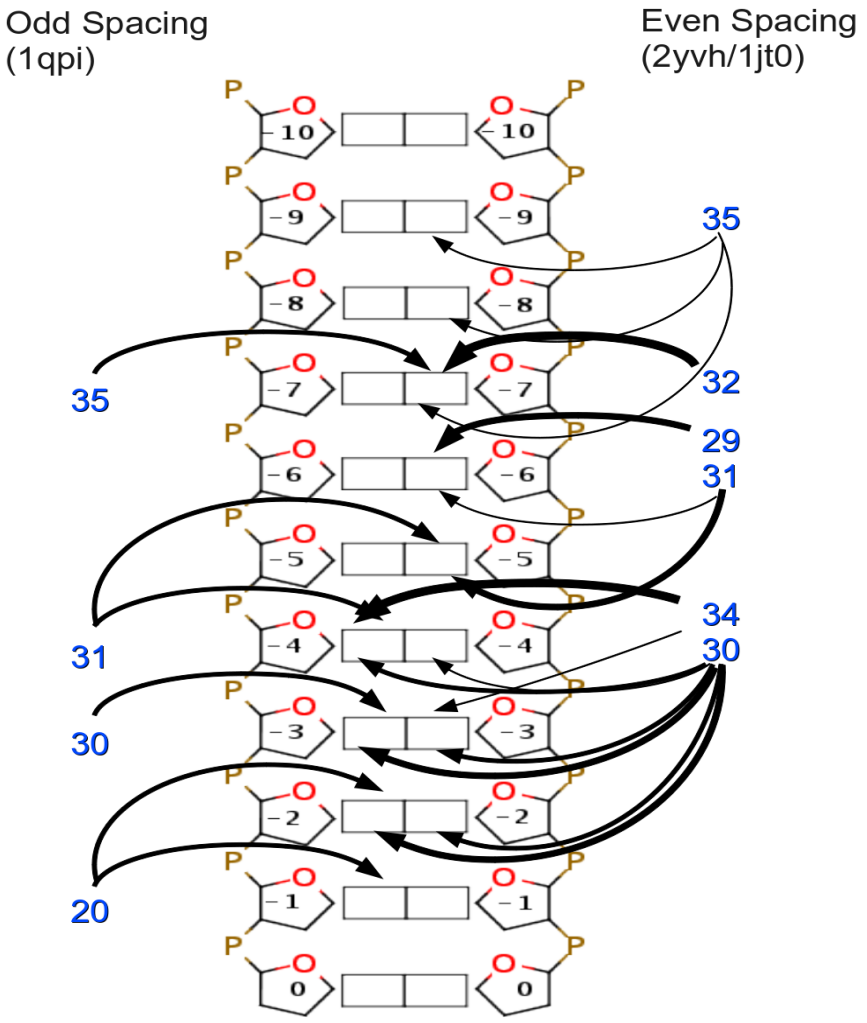
Results

To test the applicability of the motif finding method to the problem of interest, there were two tests that needed to be run. Firstly, could the motif finder determine a motif given a reasonable starting point. Secondly, could the motif finder use a predicted motif in order to determine the motif of a similar CR tag set. In the prior chapter, TetR-SPKGSYH appeared to have a gapped motif so it was chosen as the test CR tag to see if the motif finder was capable of determining a motif given a reasonable start point and also whether the hint of gapped motif was actually present in motif finding. Additionally, there was a motif that was 1 HD away, namely TetR-SGKGSYH that appeared to have a two base change in the half-site, so that was chosen as the secondary test as to whether

this method could take a similar motif as a start point and find the correct resultant motif. Since the gapped-method does not currently have a "scan mode", the correctness of the motif was judged by its visual similarity to the known motif.

In the case of the Pfam TetR family, there are three known crystal structures which were used to define the critical residues. There are two crystal structures that have an even-spacing between the motif half-site and one that has an odd-spacing between the motif half-site. The proteins were aligned and indexed based on their profile hidden markov models (profile HMM), as described in the previous chapter. Figure 3.2 shows the residues and their respective base contacts. This provides a map with which to determine the consistency of the predicted motifs with the known residue-base interactions.

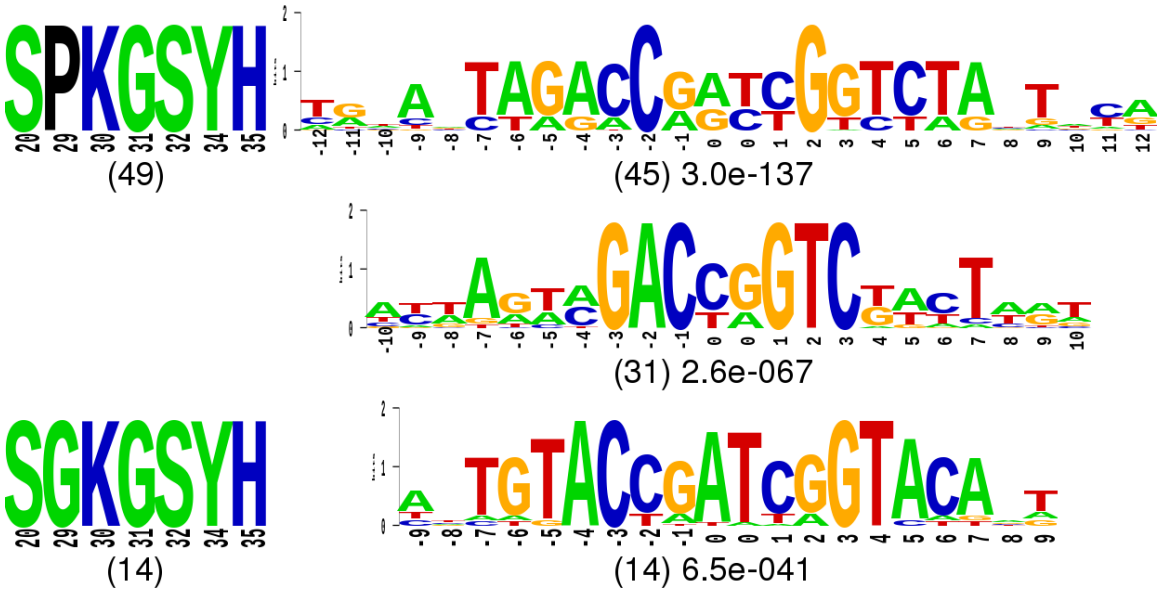
Figure 3.2. TetR residue-base contact diagram. The diagram shows the residues in blue that contact the DNA bases. The weight of the arrows pointing from the residue to the base is proportional to the number of interactions that were seen in the crystal-DNA monomers. The odd spacing implies that there was an odd number of bases between the DNA half-sites and even-spacing that there was an even number of bases. The numbering was set such that the central base of an odd-length motif was indexed as 0 and the two central bases of an even-length motif were indexed as 0.



TetR-SPKGSYH was chosen as the test case to determine if this motif finding algorithm was functionally capable of determining the correct motif. The parameters were set to allow variable gaps between 0 and 5 bp, the gap probability was allowed to vary and the half-site orientation model was chosen to be an inverted repeat. As shown in Figure 3.3, we can see that there appears to be at least two differently spaced motifs, thus this method should be able to find a single half-site with multiple gaps.

Figure 3.3. MEME determined binding motifs for TetR-SPKGSYH and TetR-SGKGSYH. A set of predicted motifs are shown along with the CR tags. Below the CR tags is the number of USPs. Below the motifs is the number of USPs that contained that

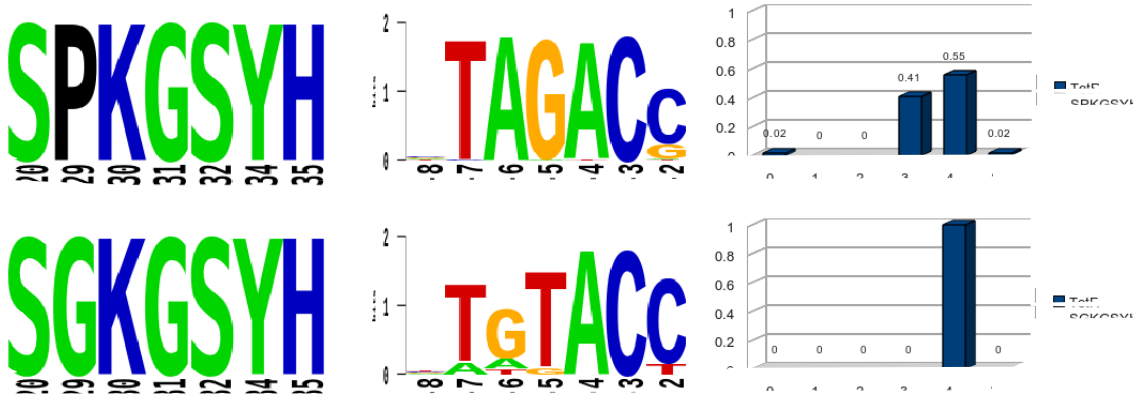
comprise the motif along with the MEME calculated e-value. The numbering is consistent with Figure 3.2.



The start-point algorithm resulted in a consensus string of NTAGACN. The results of the optimization procedure are shown in Figure 3.4. As we can see, the consensus half-site is cTAGACC, which is consistent with the MEME results, although the motif appears to be more conserved. This could be due to the fact that the original motif was a mixture of multiple different gap sizes, as evidenced by the gap probabilities. In contrast to what was expected from the prior study, only a small fraction of the half-sites appear to have a gap of 0; this could be for two reasons. Firstly, the 0 gapped motifs may not have been found given the more stringent half-site binding model. Secondly, the model was only a ZOOPS model, and thus there may have been multiple modes of binding within each USP, which would remain unaccounted for, as the method would likely have selected the most consistent set of binding sites. The size 3 gap motifs is not unexpected, as this was also seen in TetR-NPKGSYH (data not shown).

In order to create the appropriate start point for TetR-SGKGSYH, the resulting PWM from the TetR-SPKGSYH was modified as discussed in the method section by adding a pseudocount and re-normalizing. The parameters were again set to find an inverted repeat variably gapped motif with a gap size between 0 and 5 bp. The motif is shown in Figure 3.4. Interestingly in this case, there was a single unanimous gap size of 4 that was found in the dataset.

Figure 3.4. EM algorithm results for TetR-SPKGSYH and TetR-SGKGSYH. The panel on the left shows the critical residue tag along with the HMM number to correlate to Figure 3.2. The middle panels show the predicted motif half-sites, the numbering is based on an gap of 4. The panel on the right shows the chart of the final calculated gap probabilities.



Discussion

In this chapter, a method was described that would allow for gapped motif finding in the context of a bipartite motif model. This method was implemented as a two step process, first to determine a set of start points and a second to optimize this set of start points. The importance of keeping these two steps of the algorithm separate was so that the second part of the algorithm could be later applied to extending the motif predictions using a variant of the start-point determining algorithm. In the context of extending the

TF prediction space, the set of all current predictions that are a certain distance away from the CR tag of the set of proteins to be predicted are collected. In this case, a CR tag that was HD=1 away was chosen. These sets of motifs can be aligned and combined into a start point for the optimization component of the algorithm. The benefit of multiple similar motifs is that if the new motif is also variable in the same set of positions, then the process of merging the priors should actually downweight the relevant positions that will change in the motif, thus allowing the EM algorithm to have an easier time determining the solution.

The method has not been applied to the full dataset because it lacks a few modifications that would greatly improve the sensitivity and specificity of motif finding. Primarily, there are two points for improvement. Firstly, the algorithm for defining the start points is somewhat rudimentary. It is capable of finding overrepresented motifs, but is still thrown askew in some cases where there are repeats. Some sort of a complexity filter on putative motifs is likely to be needed in order to advance this method further. When determining the start points, there is some knowledge gained about the distribution of gap sizes and orientations. This knowledge is currently not being transferred to the subsequent motif finding protocol. An additional step to combine similar string motifs is likely to be beneficial in the process of both defining the start point as a more continuous quantity and also might alleviate some of these repeat-biases. In addition, the prior knowledge of the gap size and orientation is likely to be more accurate if the starting motif is also more accurate. Secondly, the EM algorithm currently is using a ZOOPS type model which allows zero or one motifs per sequence, regardless of the number of sequence fragments. However, given multiple fragments that comprise a sequence, it is

likely that more than one contains a motif occurrence, thus a ZOMPS or zero or more motifs per sequence is a superior model of the distribution of the motif. This requires additional parameters assessing the number of total expected motifs. Again, with a more robust starting point, a more accurate prior can be assessed regarding the likely number of motifs that are present in the dataset.

Additionally, the method could be improved by inclusion of a more robust statistical test of the significance of obtaining a result. Given the nature of gapped motif finding and the inclusion of multiple different orientations, there was not an obvious extension of the common methods for assessing p-values on a motif dataset. The EM algorithm is guaranteed to converge, however the resulting solution is not necessarily the optimal or a reasonable. The importance of including methods to assess the significance of a motif includes the ability to determine the statistical likelihood of such a motif appearing in the dataset by chance, as well as a method to determine the relative significance of motifs that were started at different points. The ability for relative comparison could enable the use of multiple start points with a selection step to choose the best motif at the end of the optimization procedures.

References

Bi,C. et al. (2008) A comparative study on computational two-block motif detection: algorithms and applications. *Mol. Pharm*, **5**, 3-16.

Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res*, **32**, 4979-4991.

Cardon,L.R. and Stormo,G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J.*

Mol. Biol, **223**, 159-170.

Eraso,J.M. and Kaplan,S. (2009) Half-Site DNA Sequence and Spacing Length Contributions to PrrA Binding to PrrA Site 2 of RSP3361 in Rhodobacter sphaeroides 2.4.1. *J. Bacteriol.*, **191**, 4353-4364.

Favorov,A.V. et al. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, **21**, 2240 -2245.

GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608-621.

Liu,X. et al. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-138.

Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, **8**, 269-278.

Tompa,M. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, **23**, 137-144.

Chapter 4: Progress and future directions

Current progress

In this thesis, a novel pipeline to determine TFBS in prokaryotic genomes was presented. This method is based on two basic assumptions, first that TFs have a localized regulatory structure and bind near the gene that encodes them, and second that TFs with similar DNA-interacting residues, critical residues, should have similar binding sites. As discussed in chapter 2, an initial pipeline was designed that extracted genomic data from the NCBI database, defined TF families based on their Pfam HTH signatures, used the relevant PDB structures to determine the critical residues, clustered TFs with similar critical residues and determined the putative promoter binding regions for these TFs. These putative clustered binding regions were then further processed to remove sequencing or phylogenetic biases and input into a motif finder, namely MEME. The resulting motifs were validated against RegulonDB and showed an ~80% sensitivity of prediction on both the TetR and LacI families. There were some unresolved issues, including the inability of MEME to allow for variably gapped binding sites and to allow for differing binding orientations as well.

In order to address these issues, an EM-based variable gapped bipartite-model motif finder along with a rudimentary enumerative equivalent was developed as described in chapter 3. Conceptually, this motif finder was designed to not only find motifs in the large clusters as in chapter 2, but also extending the current set of predictions via a nearest-neighbor approach. As shown on a set of TF examples from the TetR HTH family, this method appears to be functional, but as discussed in chapter 3, there still remains some work to be done before the method can be broadly applied and integrated into the bioinformatics pipeline. In this chapter, some additional extensions to

the methods will be described along with their potential broader implications in the scientific field.

Expanding family repertoire

The method that has been outlined should be easily extensible to other HTH families. The bioinformatics pipeline is structured to be flexible in this regard, however, application may require more careful thought. To this end, the pipeline described in chapter 2 was extended with most of the steps becoming nearly fully-automated and the addition of a configuration file simplifies porting the code to other systems or users. As with any large scale project, there are occasional issues that arise due to evolving file formats or data sources that do not conform to the standard specifications, however these are generally flagged by the pipeline for further inspection. Using the additional automation, the pipeline can be directly applied to all of the remaining Pfam families with ~50% overall success rate, however, the success rate on many families is quite low and these shortcomings need to be explored further. For example, in the AraC family, many proteins contain multiple putative DNA binding domains, but the binding modality of these domains does not appear to be conserved (Martin and Rosner, 2001). There are family members that contain two domains that bind DNA (Rhee et al., 1998), whereas other family members only contain a single domain. In some cases, there are family members with multiple domains, but only one actually binds the DNA (Kwon et al., 2000). Also in AraC, a subset of proteins binds via direct and/or indirect repeats (Carra and Schleif, 1993; Gallegos et al., 1997). Given these sorts of problems, there are some Pfam families that may require special attention rather than simply applying the default analysis pathway. The novel motif finder that was described in chapter 3 may help in this

matter, as it provides a flexible platform on which to develop a more robust pipeline. The extensions described in the discussion section of that chapter would be important first steps in order to analyze additional motif families.

Currently, the method assumes that all proteins in a family are using the same sets of critical residues in order to interact with the DNA. This assumption may not be correct in all cases, as the critical residue set was assessed by taking the union of all possible protein-DNA contacts. Generally this is a valid assumption, but in some cases, there may be a specific subset of the critical residues that are used in determining DNA specificity in specific HTH proteins (Wintjens and Rooman, 1996). Normally, the consequence of the faulty assumption will simply lead to smaller cluster sizes and fewer motif predictions, however it could also lead to motif predictions that do not represent the true diversity of the proteins in the cluster. The challenges that would arise from such an analysis would be two-fold. First, a method would need to be built such that it could determine which subset of residues are being used by a particular TF. Second, discrepancies based on the first analysis would need to be resolved; if the method predicts that a specific TF was able to use multiple different subsets of critical residues and the results of motif finding are inconsistent this would need to be resolved.

Many ongoing sequencing projects are generating metagenomic sequence data. The pipeline was theoretically built to handle such datasets, and has placeholders in many of the routines to include metagenomic datasets, but as of yet has not been fully tested using metagenomic sequences. In the context of the human microbiome project and other sequencing projects, this will become an increasingly important source of information and it will be imperative to include metagenomic sequence information into the pipeline.

The potential problems with including such data is the length of the contigs and the relative position of the TF or the relevant promoters/operons. If the TF is located too close to either end of the contig or the contig is too short, then it is likely that the upstream promoter, downstream promoter and autoregulatory promoter will not all be present. This would affect the assumption of each concatenated USP containing a motif. Technically, this could also be an issue in WGS datasets as well, but it has not proven to be too much of a problem. Additionally, given the context of metagenomic sequencing, these contigs could potentially represent chimeras of multiple different organisms that were assembled into a single contig, thus this could also throw a motif finder askew. To solve the first problem, a simple data filtration criteria could be assessed such that if a contig does not contain at least 2 of the 3 relevant promoters, it is not included in the motif finding dataset. Additionally, to solve both problems, the motif finder from chapter 3 could be easily extended such that it is capable of relative sequence weighting and a position specific prior. If the confidence in the metagenomic sequencing project and the resulting contigs therein, is not high, this could be reflected in a lower relative sequence weight which would effectively decrease its impact on the final motif. The exact relative weighting of the metagenomic sequencing is not entirely clear without some testing, and would be an important step in analyzing such data.

Extending motif predictions

As discussed in chapter 3, there are three major improvements that need to be implemented in the current gapped motif finder before it can be deployed in a larger scale. Firstly, the start-point algorithm needs to be refined. It needs to be more resilient to repeat regions and in some ways, clustering the top set of (l,d)-mer matches might

provide for a more accurate start point. The start-point definition might also improve if there were a complexity filter on the set even before clustering. Given superior start-point definitions, more accurate priors could be assessed that would be input into the optimization procedure. Secondly, improvements in the optimization part of the algorithm would help to more accurately model the expected distribution of the motifs as ZOMPs as opposed to ZOOPs. Finally, there needs to be a rigorous analysis of the significance of a motif. Since EM is guaranteed to converge, the statistical importance of a motif is much more important than if motif finding was done using a method that would not converge unless a reasonably significant motif was present. This significance value could be used to not only assess the degree of certainty that the motif is valid, but also to compare different motif predictions to determine the one that is most consistent with the data.

One of the main driving factors for developing a motif finder was in order to have a method that was capable of optimizing a motif from a start point in the context of extending known motifs predictions to similar clusters. In each Pfam family, there are many motif predictions, however these predictions are not complete, as when there are fewer than 10 USPs, no predictions are made. This minimum limit of 10 USPs is somewhat arbitrary, however, if there are too few sequences provided to a motif finder then the motif finding may be unstable. As discussed in chapter 2, we can see that the idea of using similar CR tags as defined by a HD criteria in order to predict the binding specificity of additional motifs would allow for a significant increase in the number of predicted members of a Pfam family. This extension would require a few additional modifications to the current procedure in order to work. Firstly, the motifs of the similar CR tags would need to be converted into a form that is conducive to the motif

optimization procedure. This could be as simple as inputting a slightly muted version of the motif by adding a global pseudocount and re-normalizing as discussed in chapter 3. However, in the case where there are more than one CR tags that are within the specified similarity criteria, a more robust method of aligning the motifs likely using a modified variant of STAMP (Mahony and Benos, 2007) and trimming the resulting familial binding profile (FBP) is likely to be effective. The modification of STAMP included outputting the intermediate full alignments as opposed to simply the consensus representations. The benefit of STAMP is that one can weight the individual motifs that are input into the procedure, so the confidence in the prediction is also reflected in this stage of the analysis as well. The pipeline has been built with the concept of "rounds" or iterations, and thus one could extend these predictions while tracking the level of extrapolation that has occurred. In round 1, it could be a single level of extrapolation, such that only CR tags with HD=1 are used to extend the predictions, whereas in round 2, HD=2 or extrapolated HD=1could be used to further extend the predictions. Keeping the primary data and the extrapolations separated would allow for easier interpretation of the data in further analyses, potentially require assessment of the confidence. The separation of the datasets would also assist in tracing of any potential problems in the extrapolative process.

Regulatory Code

One such analysis could include defining a regulatory code for the TF families of interest. The regulatory code is an elusive "function" by which if the protein sequence is known, the DNA binding site can be predicted, similar to how with the nucleic acid coding sequence, the protein sequence can be inferred. Generic deterministic versions of

this regulatory code that apply to all DNA-binding proteins do not appear to exist (Matthews, 1988). However, family specific codes can be elucidated (Benos et al., 2002; Noyes et al., 2008). In this context, there are multiple transcription factor families of the HTH class, and thus might be amenable to such a regulatory code. The importance of a regulatory code in these contexts is that it would enable motif predictions of CR tags that either are not amenable to discovery by the current/future protocols, or that do not even exist in the dataset. If these predictions can be made, this might be quite useful in the synthetic biology arena, where custom TF-DNA interactions would be a key part of designing novel transcriptional networks.

Methods to elucidate the regulatory code could include a mutual information analysis (Mahony et al., 2007). The drawback to the mutual information method is that it is quite stringent on quality the input dataset, thus a validated set of data would need to be input. This is still not ideal though, as mutual information methods normally require a large amount of data in order to draw their conclusions and the validation process would likely eliminate some lower quality input points. A simpler method to determine a correlation between the residue and the DNA might be to take the set of motifs whose CR tags are 1 HD away from each other and calculate a column-based distance metric between their motifs. The columns with the furthest distance are likely the ones that are interacting with the residue that has changed. The complication in this method is that the motif half-sites need to be aligned in a consistent manner so that the resulting correlations are consistent and can be compiled across multiple comparisons, maybe using STAMP in order to align the half-sites. Once the correlations between DNA position and residue number have been determined, then a closer inspection of the co-variance can be

undertaken to see if there is a residue-based code underlying the motif changes. Again, this method will require a significant corpus of data, however, as discussed in chapter 3, the new motif finder should be able to extend the predictions to the HD=1 CR tags successfully, thus providing the additional needed data.

Informing Biophysical models

There has been some work on generating biophysical models of TF binding to DNA, however, these studies have been plagued by a paucity of known TF-crystal structures. Using this pipeline, there will be a large dataset correlating transcription factor protein sequences to their potential DNA binding sites. This dataset could be used to provide insight into the specific mechanisms underlying transcription factor-DNA binding interactions. An easy way to include the data would be as a larger test set to determine which of the current methods is best at predicting HTH binding specificities *de novo*. Alternatively, the same results from the testing procedure could be used to pinpoint problems in the current prediction methods. If a method is capable of predicting a motif, but not its HD=1 correlate, than that could provide useful information about a potential problem in the parameters describing a specific residue-DNA interaction pair. However, the outlined method will not be able to help in terms of the lack of explicit waters and difficulty in accurately calculating electrostatic interactions for the highly charged DNA.

Alternatively, the large corpus of data could be used to calculate “fudge-factors” or correction terms that would allow for a semi-empirical approach to biophysical binding site prediction. In this hybrid procedure, first the binding site energy would be calculated using an all-atoms model, and then depending on the protein-DNA interactions that are likely to be present given the regulatory code analysis, certain correction terms

could be applied. These would allow for some leeway in the prediction accuracy, such that it was not entirely dependent on the force-field based terms. Again, the simplest method would be to apply these additional terms during the ranking procedure to determine whether a sequence is likely to be bound by the DNA or not. The relative weight of the correction terms versus the computed force-field energy would likely depend on several factors including the force field used, the optimization procedure employed and the confidence in the predictions.

Synthetic biology

The method that has been described will provide a large dataset of TF-DNA interactions. Assuming the successful extension of the method to a large number of TF families, many of the current methods could be applied to generate a relatively robust transcriptional regulatory network. Even though the method has only been applied/tested on HTH families, these comprise ~80% of *E. coli* TFs. Thus the resulting regulatory network should be relatively complete.

With a transcriptional regulatory network in hand as well as the predicted TF-DNA interactions in bacterial systems, it becomes more feasible to manipulate the cellular transcriptional network. There has been increased interest in developing synthetic organisms or modifying current organisms to perform specific tasks or functions, known as synthetic biology (Purnick and Weiss, 2009). Recently a minimal organism has been generated that could be a scaffold for future research (Gibson et al., 2010). Additionally, there has been significant interest in developing organisms that can be used in the production of biofuels or for bioremediation (Löffler and Edwards, 2006; Lovley, 2003;

Weber et al., 2010).

One of the limiting factors in engineering organisms is the ability to determine the effect of a newly inserted gene or operon on the fitness of the cell. This is a complex problem that is not only limited to the gene regulatory network, but can include the intermediate RNA, the protein products and the resulting metabolites. In order to optimize the function of the cell, all of these factors need to be taken into account and this highly depends on what is being optimized as well, production of a certain end product or sheer biomass. The gene regulatory network is the lowest level at which this optimization can happen, so it is likely to have a large impact on the results. Specifically, given a bacterial simulator, that takes into account the factors that are important for growth and production of metabolites/products, it is easy to envision how modulating the affinity of the TFs in the system for their respective targets could allow for the optimization of production of a certain metabolite, by simply downweighting (or substituting with suboptimal binding sites) non-required pathways and upweighting (or substituting with optimal binding sites) for the pathways of interest. This would need to be done in a careful manner, as there may be non-linear dependencies that underly these pathways, but the gene regulatory network of an organism along with the binding specificities of the component TFs would provide an invaluable resource to the synthetic biology community. Additionally, this could be done for a wide variety of organisms, as the method has been applied to all currently sequenced organisms.

References

Benos,P.V. et al. (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. .

Bioessays, **24**, 466-475.

Carra,J.H. and Schleif,R.F. (1993) Variation of half-site organization and DNA looping by AraC protein. *EMBO J*, **12**, 35-44.

Gallegos,M.T. et al. (1997) Arac/XylS family of transcriptional regulators. *Microbiol. Mol. Biol. Rev*, **61**, 393-410.

Gibson,D.G. et al. (2010) Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science*, **329**, 52-56.

Kwon,H.J. et al. (2000) Crystal structure of the Escherichia coli Rob transcription factor in complex with DNA. *Nat. Struct. Biol*, **7**, 424-430.

Löffler,F.E. and Edwards,E.A. (2006) Harnessing microbial activities for environmental cleanup. *Curr. Opin. Biotechnol*, **17**, 274-284.

Lovley,D.R. (2003) Cleaning up with genomics: applying molecular biology to bioremediation. *Nat. Rev. Microbiol*, **1**, 35-44.

Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, **35**, W253-W258.

Mahony,S. et al. (2007) Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, i297-304.

Martin,R.G. and Rosner,J.L. (2001) The AraC transcriptional activators. *Curr. Opin. Microbiol*, **4**, 132-137.

Matthews,B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294-295.

Noyes,M.B. et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277-1289.

Purnick,P.E.M. and Weiss,R. (2009) The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol*, **10**, 410-422.

Rhee,S. et al. (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl. Acad. Sci. U.S.A*, **95**, 10413-10418.

Weber,C. et al. (2010) Trends and challenges in the microbial production of lignocellulosic bioalcohol fuels. *Appl. Microbiol. Biotechnol*, **87**, 1303-1315.

Wintjens,R. and Rooman,M. (1996) Structural Classification of HTH DNA-binding Domains and Protein - DNA Interaction Modes. *Journal of Molecular Biology*, **262**, 294-313.

Appendix 1: Regulation of the *Drosophila Enhancer of split* and *invected-engrailed* Gene Complexes by Sister Chromatid Cohesion Proteins[2]

2 This chapter was adapted from: Schaaf, C. A., Misulovin, Z., **Sahota, G.**, Siddiqui, A. M., Schwartz, Y. B., Kahn, T. G., Pirrotta, V., Gause, M. & Dorsett, D. Regulation of the Drosophila Enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins. *PLoS One* **4**, e6202 (2009). I created scripts to re-analyze the cohesin ChIP-chip data in order to define cohesin (Nipped-B/Smc1) +/- PolII binding in a rigorous manner and helped correlate the microarray data with these definitions.

Abstract

The cohesin protein complex was first recognized for holding sister chromatids together and ensuring proper chromosome segregation. Cohesin also regulates gene expression, but the mechanisms are unknown. Cohesin associates preferentially with active genes, and is generally absent from regions in which histone H3 is methylated by the Enhancer of zeste [E(z)] Polycomb group silencing protein. Here we show that transcription is hypersensitive to cohesin levels in two exceptional cases where cohesin and the E(z)-mediated histone methylation simultaneously coat the entire *Enhancer of split* and *invected-engrailed* gene complexes in cells derived from *Drosophila* central nervous system. These gene complexes are modestly transcribed, and produce seven of the twelve transcripts that increase the most with cohesin knockdown genome-wide. Cohesin mutations alter eye development in the same manner as increased *Enhancer of split* activity, suggesting that similar regulation occurs in vivo. We propose that cohesin helps restrain transcription of these gene complexes, and that deregulation of similarly cohesin-hypersensitive genes may underlie developmental deficits in Cornelia de Lange syndrome.

Introduction

The cohesin protein complex holds sister chromatids together, ensuring their proper segregation upon cell division [1-3]. Cohesin has a ring-like structure that encircles DNA [4,5], formed by the Smc1, Smc3, Rad21 and Stromalin (SA) proteins. In most organisms, cohesin binds chromosomes throughout interphase, and several findings indicate that it regulates gene expression. The *Drosophila* Nipped-B protein that loads cohesin onto chromosomes facilitates activation of the *cut* and *Ultrabithorax* homeobox genes, and cohesin inhibits *cut* expression [6-9]. *Drosophila* cohesin facilitates expression of a steroid hormone receptor and axon pruning in non-dividing neurons [10,11], and the Rad21 cohesin subunit encoded by *verthandi* (*vtd*), was identified genetically by its opposing effect to Polycomb group (PcG) silencing of homeotic genes [12,13]. Rad21 also facilitates expression of zebrafish Runx genes in a cell-type specific manner [14].

To understand how Nipped-B and cohesin regulate gene expression, their binding was mapped in the genomes of *Drosophila* cultured cells, revealing that they co-localize genome-wide [15]. Cohesin was also mapped in the human genome [16], and in 3% of the mouse genome [17]. All three studies show that cohesin binds many genes, and that binding is particularly enriched around transcription start sites.

In mammals, cohesin co-localizes extensively with the CCCTC-binding factor (CTCF) that functions as a transcriptional insulator, and cohesin contributes to insulation [16,17]. CTCF is thought to function by forming long-range chromosome loops, and cohesin and CTCF support transcription-dependent loops in the human apoliporotein

gene cluster [18] and a developmentally-regulated loop at the *IFNG* cytokine locus in mammalian T cells [19].

There are also links between insulators and cohesin in *Drosophila*. A 75 kb domain of cohesin that covers the active *Abd-B* gene in the bithorax complex is flanked by a CTCF site near the 5' end of *Abd-B*, and the Fab-7 insulator downstream of *Abd-B* [15, 20], suggesting that insulators define some cohesin domains. On the basis of genetic evidence it was suggested that cohesin blocks enhancer-promoter interactions in *cut*, and that Nipped-B counters this insulation by controlling cohesin binding [8]. Most recently, genome-wide mapping revealed that the *Drosophila* CP190 insulator protein co-localizes extensively with cohesin [21].

Many differences in cohesin binding between different *Drosophila* cell lines correlate with differences in transcription, with cohesin binding a gene only in those cells in which the gene is active [15]. Cohesin extensively overlaps RNA polymerase II (PolII) genome-wide, but is almost always absent from regions in which the E(z) protein of the PRC2 PcG silencing protein methylates histone H3 on the lysine 27 residue (H3K27Me3).

There are rare cases where cohesin overlaps H3K27Me3 over large regions in ML-DmBG3 (BG3) cells [22] derived from *Drosophila* central nervous system. One of these is the *Enhancer of split* complex [E(spl)-C] that contains twelve genes, including seven basic helix-loop-helix (bHLH) genes that repress neural fate [23]. Another is the *invected-engrailed* complex with two homeobox genes expressed in posterior developmental compartments [24-26]. The unusual pattern prompted us to determine if cohesin regulates these gene complexes. We find that genes in these complexes are

expressed at modest levels, and that in sharp contrast to most cohesin-binding genes, reducing Nipped-B or cohesin levels dramatically increases their transcription.

Results

Cohesin and RNA polymerase II (PolII) binding overlap extensively genome-wide, while cohesin shows a negative correlation with the H3K27Me3 mark made by the PRC2 PcG silencing complex [15]. PcG target genes such as *Abd-B* or *cut* bind little or no cohesin in cells in which they are silenced, but bind cohesin over large regions of 75 and 150 kb in cells in which they are transcribed [15].

While comparing the cohesin and H3K27Me3 patterns, we noted eight unusual regions of extensive overlap ranging in length from 4.8 to 80.9 kb in the genome of BG3 cells derived from central nervous system, and only two such regions in Sg4 cells of embryonic origin (Table A1.S1). Strikingly, two of the BG3-specific overlaps align perfectly with developmentally-important gene complexes. Figure A1.1 shows the association of cohesin, RNA polymerase II (PolII), and H3K27Me3 with the *Enhancer of split* and *invected-engrailed* complexes in BG3 and Sg4 cells. In BG3 cells, the 50 kb length of the E(spl)-C binds cohesin and has extended regions of H3K27Me3. Six genes (*HLHmδ*, *HLHmβ*, *mα*, *HLHm3*, *HLHm7*) bind PolII. By contrast, in Sg4 cells, only three E(spl)-C genes bind cohesin (*HLHmβ*, *HLHm3*, *m6*), six bind PolII (*HLHmδ*, *HLHmβ*, *m2*, *HLHm3*, *m6*, *HLHm7*), and there is no H3K27Me3. Similar to the E(spl)-C, the *invected-engrailed* complex is also coated by cohesin, and has extensive H3K27Me3 in BG3 cells (Figure A1.1). The cohesin domain extends from upstream of the *invected*

transcription start site to a region upstream of *engrailed* that contains a Polycomb Response Element (PRE) and sequences required for interactions with transcriptional enhancers [27]. The H3K27Me3 region also starts upstream of *invected*, but extends 50 kb past the PRE, over a region that regulates *engrailed* [28]. In Sg4 cells, H3K27Me3 also coats the *invected-engrailed* complex and the regulatory region, but there is no PolII and little cohesin, as is typical for PcG-targeted genes [15].

Cohesin Regulates the E(spl)-C and *invected-engrailed* Complex in BG3 Cells

The unusual cell-type specific overlap of cohesin and H3K27Me3 that covers the E(spl)-C and *invected-engrailed* raised the possibility that cohesin might regulate their expression. Genome-wide, 480 genes have H3K27Me3 ($p \leq 10^{-3}$) in their transcribed regions in BG3 cells, and only 64 (13%) of these bind PolII, including the genes in the E(spl)-C and the *invected-engrailed* complex. Although PcG proteins bind PREs of some target genes in both the inactive and active states, for the genes examined, H3K27Me3 covers the transcribed region only when they are silent [29-32]. We measured transcripts to compare expression of the E(spl)-C and *invected-engrailed* complex in BG3 and Sg4 cells. Consistent with the binding of PolII, seven E(spl)-C genes (*HLHmδ*, *HLHmγ*, *HLHmβ*, *mα*, *m2*, *HLHm3*, *HLHm7*), *invected*, and *engrailed* are transcribed in BG3 cells (Figure A1.2A). An overlapping set of six E(spl)-C genes (*HLHmδ*, *HLHmβ*, *mα*, *m2*, *HLHm3*, *m6*) are expressed in Sg4 cells at levels similar to those seen in BG3 cells (Figure A1.2A), but *invected* and *engrailed* are essentially silent. Thus at the *invected-engrailed* complex, which is coated by H3K27Me3 in both cell types, the presence of

Nipped-B and cohesin correlates with expression, suggesting that cohesin prevents complete silencing, and/or that incomplete silencing promotes cohesin binding.

We used RNAi to knock down Nipped-B and cohesin to see if this alters expression of the *Enhancer of split* and *invected-engrailed* complexes. Knockdown of Nipped-B had little effect on cohesin levels, while Rad21 knockdown slightly reduced SA as previously noted [33], and SA RNAi reduced Rad21 (Figure A1.2E). SA and Rad21 interact, making it likely that they stabilize each other. In several experiments with BG3 cells, knockdown of Nipped-B, Rad21 or SA was maximal within two days, and on the order of 80% for several days (Figure A1.2B,E). Knockdown in Sg4 cells was maximally 60% after two successive treatments.

We saw large increases in E(spl)-C, *invected* and *engrailed* transcripts in BG3 cells six days after Rad21, Nipped-B or SA RNAi in all of several experiments (Figure A1.2B,C). The increases varied somewhat between experiments. In Figure A1.2C, the *HLHmδ* transcripts increase 130-fold by day 6 in one experiment, and 25-fold in another with Rad21 RNAi, representing some of the largest and smallest increases observed in the nearly forty independent Rad21 RNAi experiments that were performed. Within each experiment using the same cell passage, however, effects were similar between Rad21 and Nipped-B knockdown, or between Rad21 and SA RNAi (Figure A1.2C). Thus we attribute the variability in the fold-effects from experiment to experiment to unknown differences in the physiology or growth state of the cells between passages, and conclude that overall, Nipped-B and cohesin have similar effects on gene expression. We measured transcripts up to 13 days after RNAi, when Nipped-B (not shown) or Rad21 (Figure A1.2B) recover. The E(spl)-C and *invected-engrailed* transcripts start to decrease, but are

still above initial levels (Figure A1.2B).

Nipped-B or cohesin RNAi had little effect on expression of the E(spl)-C in Sg4 cells (Figure A1.2D), including the cohesin-binding *HLHm3* and *m6* genes. There was also no effect on the silenced *invected* and *engrailed* genes. Although Rad21 and Nipped-B knockdown was less efficient in Sg4 cells (Figure A1.2E), as shown below, Rad21 knockdown of 30 to 50% in BG3 cells alters E(spl)-C RNA levels. We conclude that the E(spl)-C and *invected-engrailed* are less sensitive to cohesin dosage in Sg4 than in BG3 cells, as might be expected from the substantial differences in cohesin binding between the two cell types.

On day 3 after Nipped-B RNAi, some E(spl)-C transcripts (*HLHmγ*, *mα*, *m2*, *HLHm3*) decrease (Figure A1.3), yet show large increases by day 6 (Figure A1.2). Similar decreases at day 3 were seen in all Nipped-B RNAi experiments. To see if a biphasic effect also occurs with Rad21, we used different amounts of dsRNA to control RNAi efficiency. A 30% knockdown decreased most E(spl)-C transcripts, while a 55% reduction decreased some and increased others (Figure A1.3). Thus Rad21 has a biphasic effect similar to Nipped-B.

E(spl)-C transcripts are miRNA targets [34], and we considered the possibility that cohesin knockdown decreases miRNA activity to increase transcript stability in BG3 cells. Rad21 knockdown, however, had little effect on the stability of E(spl)-C transcripts (Table A1.S2), and we therefore conclude that cohesin RNAi elevates E(spl)-C transcription.

Nipped-B or Rad21 knockdown slowed but did not arrest cell division in BG3 or Sg4 cells, consistent with previous findings in *Drosophila* cells [33]. Sister chromatid

separation increased 2 to 3-fold over controls, but there was no increase in hyperploid cells, indicating that the minor cohesion deficits did not affect segregation (Table A1.S3). Nipped-B or cohesin RNAi did not increase cell death, as determined by trypan blue staining.

Polycomb Represses the E(spl)-C in BG3 Cells

In contrast to *engrailed*, the E(spl)-C has not previously been reported to be a PcG target. We used RNAi knockdown of the Polycomb (Pc) subunit of the PRC1 complex to see if PcG proteins repress the E(spl)-C in BG3 cells. With a Pc knockdown of some 70%, most E(spl)-C transcripts increased several-fold by day 6, indicating that in addition to cohesin, PRC1 restrains their expression (Figure A1.4). The *invected* and *engrailed* RNA levels did not change (Figure A1.4), although *Abd-B*, which is PcG-silenced and does not bind cohesin [15], showed up to 1200-fold increases in transcript levels with Pc knockdown (not shown). The lack of effects on *invected* and *engrailed* transcripts suggests that Pc is not strongly limiting for their repression in BG3 cells. Pc is only weakly limiting for repression of *engrailed* in embryos, and is less limiting than other PcG proteins for repression of many target genes in imaginal discs [31,35].

The CP190 Insulator Protein Does Not Regulate E(spl)-C and *invected-engrailed* Transcription in BG3 Cells

Cohesin can regulate gene expression by contributing to activity of the CTCF insulator protein and insulator-mediated looping in mammalian cells [16-19]. Drosophila

has many insulator proteins, including CTCF, Su(Hw), GAF, and BEAF. All co-localize extensively genome-wide with the CP190 protein, which is required for CTCF and Su(Hw), and likely also for GAF and BEAF insulator activities [21,36]. We used RNAi to knockdown CP190 protein by approximately 90%, but there was little or no change in the level of E(spl)-C and *invected-engrailed* transcripts six days after RNAi treatment (Figure A1.5). Rad21 knockdown substantially increased E(spl)-C and *invected-engrailed* transcripts in the same experiment. CP190 knockdown also had no significant effect four or eight days after RNAi treatment (not shown). These results argue that the effects of Nipped-B and cohesin on transcription of these gene complexes, which are substantially larger than the effects of cohesin on insulator function seen in mammalian cells, are unlikely to result from changes in insulator function.

Nipped-B and *Rad21* Mutations Alter *Notch^{split}* Mutant Phenotypes

We used mutant phenotypes of the *split* missense mutation in the *Notch* receptor gene ($N^{spl\text{-}1}$) that are sensitive to E(spl)-C activity to test if cohesin regulates the E(spl)-C in vivo. $N^{spl\text{-}1}$ reduces activation of proneural genes, thereby decreasing the number of photoreceptors in the eye, and altering bristles [37]. E(spl)-C duplications, the $E(spl)^{D}$ gain-of-function allele, and forced overexpression of some E(spl)-C genes increase the severity of the eye phenotype [37-40], while E(spl)-C deletions suppress [41].

We tested if two loss-of-function *Rad21* mutations [12], the vtd^{36} missense mutation, and the $vtd^{y26\text{-}6}$ splice site mutation, dominantly alter the $N^{spl\text{-}1}$ mutant phenotypes. Both increased the severity of the eye phenotype, and consistent with a previous report [9], the $Nipped\text{-}B^{407}$ null allele suppressed the eye phenotype (Figure A1.6). Both *Rad21* alleles also decreased the number of scutellar macrochaete (Figure

A1.6). The simplest explanation is that reduced Rad21 dosage increases E(spl)-C expression in the developing eye and bristles, reducing the number of cells that adopt neural fate and become photoreceptors or bristles.

Knockdown of either Nipped-B or Rad21 increases E(spl)-C transcription in BG3 cells. Thus the opposing effects of *Nipped-B* and *Rad21* mutations on the N^{spl-1} eye phenotype appear contradictory. We posit, however, that they reflect biphasic effects on E(spl)-C expression similar to those seen in BG3 cells (Figure A1.3). Heterozygous *Nipped-B* null mutations reduce *Nipped-B* mRNA by only 25% in vivo [8] and thus their suppression of N^{spl-1} could reflect a decrease in E(spl)-C transcription caused by a biphasic effect. Although the biphasic effect is transitory with an 80% Nipped-B reduction in BG3 cells, it may last longer with a 25% reduction in vivo, and the critical phase for E(spl)-C expression in the developing eye at the morphogenetic furrow likely lasts for a much shorter time than three days [42].

Cohesin's Effects on E(spl)-C and *invected-engrailed* Transcription in BG3 Cells are Exceptional

We measured effects of Nipped-B and Rad21 on gene expression in BG3 cells using microarrays to (a) see if the effects of cohesin on E(spl)-C and *invected-engrailed* expression are unique, (b) look for effects of cohesin on regulators of E(spl)-C and *engrailed*, and (c) obtain a comprehensive view of the role of cohesin in gene expression. We used two samples for three days after RNAi treatment, one four day and one six day sample for both Nipped-B and Rad21, and mock RNAi controls for each time point. Comparing $\log_2$ expression values, the genome-wide correlation coefficients between the four control samples were greater than 0.99.

Strikingly, seven of the twelve transcripts that increase the most six days after Rad21 RNAi treatment are from the E(spl)-C and *invected-engrailed* (Figure A1.7, Figure A1.S1, Table A1.S4). Biphasic effects are seen, as some E(spl)-C transcripts decrease after 3 days of Nipped-B RNAi, but increase by day 6 (Figure A1.S1, Table A1.S4). E(spl)-C and *invected-engrailed* transcripts are present at relatively low levels in mock RNAi controls (Figure A1.S2, Table A1.S4). Thus the E(spl)-C and *invected-engrailed* are expressed at modest levels, and are unusually sensitive to cohesin.

Other genes located in regions of cohesin-H3K27Me3 overlap also significantly increase in expression with cohesin or Nipped-B knockdown, including *jing*, *Psc*, *Su(z)2*, *hth*, and *Lim1* (Tables S1 and S4). The increases are from 1.4 to 4-fold, and less than those observed with the E(spl)-C and *invected-engrailed*, but these genes are already expressed at 10 to 500-fold higher levels than the E(spl)-C prior to cohesin or Nipped-B knockdown, despite the extensive H3K27Me3 in their transcribed regions (Table A1.S4). After knockdown, their expression ranges from 2-fold less to 4-fold more than E(spl)-C transcripts, suggesting that the lower fold-increases in expression of these genes with cohesin knockdown reflects their initial higher expression levels. We conclude that all genes in regions of substantial cohesin-H3K27Me3 overlap in BG3 cells are not silenced, and are negatively regulated by cohesin.

Cohesin Knockdown Increases Expression of Notch Pathway Genes

BG3 cells are derived from central nervous system, but the proneural genes (*ac*, *sc*, *l'sc*, *ato*, *da*) that promote E(spl)-C expression [42,43] are not expressed (Table A1.S4). E(spl)-C genes are activated by Notch, and the genes encoding Notch (N), the Suppressor of Hairless [Su(H)] protein that tethers the Notch intracellular fragment to

target genes, the Mastermind (Mam) coactivator, and both the Delta (Dl) and Serrate (Ser) Notch ligands are expressed. Cohesin RNAi increases *Ser* ligand transcripts 6-fold on day 3 and 25-fold by day 6, and thus elevated Notch signaling may help increase E(spl)-C transcription (Figure A1.S1, Table A1.S4).

Lack of proneural gene transcripts suggests that Notch, which alone is insufficient to activate E(spl)-C genes [42], cooperates with other unknown activators to induce E(spl)-C expression. Binding sites for many transcription factors are conserved in the E(spl)-C between *Drosophila* species [44] and some of these (*Adf1*, *broad*, *Trl*, *Eip74EF*, *dorsal*, *tramtrack*, *zeste*) are expressed in BG3 cells (Table A1.S4).

Effects of cohesin on the Notch pathway cannot explain the effects of *Nipped-B* and *Rad21* mutations on N^{spl-1} phenotypes described above. If *Rad21* mutations increase Notch signaling, they should increase proneural gene expression and suppress N^{spl-1}. *Nipped-B* mutations do suppress the eye phenotype, but they have little effect on the N^{spl-1} bristle phenotype, the N^{md-1} wing margin phenotype, or the N^{Ax-E2} wing vein phenotype, indicating that they do not increase Notch signaling in vivo [9]. Thus a biphasic effect on E(spl)-C transcription remains the simplest explanation for the opposite effect of *Nipped-B* and *Rad21* mutations on the N^{spl-1} eye phenotype.

Embryonic regulators of *engrailed* (*ftz*, *eve*, *prd*, *slp*, *odd*) are not expressed before or after cohesin RNAi (Table A1.S4). The genes that regulate *engrailed* in later stages, however, are unknown, and thus indirect effects of cohesin RNAi on *invected-engrailed* expression cannot be ruled out. We note, however, that the modest changes in expression seen for most genes are unlikely to cause the unusually large changes in *invected* and *engrailed* expression.

Cohesin Has Minimal Effects on PcG and trxG Genes

We considered the possibility that cohesin could regulate the E(spl)-C and *invected-engrailed* through effects on PcG or trxG gene transcription. Most of these genes, however, are not affected by cohesin RNAi (Table A1.S4). Exceptions are an increase of 80% in *Pc* transcripts and a 2-fold increase in *Psc* expression by day 6, but this should increase silencing and reduce transcription. A few trxG transcripts (*brahma*, *osa*, *ash1*, *Trl*, *Bre1*) increase less than 2-fold. Cohesin had no significant effect on any of the 394 genes with H3K27Me3 that do not bind cohesin, most of which are not detectably expressed above background levels, including all the genes in the bithorax and Antennapedia complexes (Table A1.S4).

Cohesin Directly Regulates Gene Expression

The genome-wide effects of Nipped-B and Rad21 RNAi on gene expression after six days were very similar, with a correlation between the $\log_2$ Nipped-B/control and $\log_2$ Rad21/control expression ratios of 0.93 (Figure A1.7). Thus, with very few exceptions, Nipped-B and cohesin regulate the same genes to similar extents. Genome-wide, slightly more than 10% of transcripts showed statistically significant changes in one or more RNAi treatments, with 959 transcripts increasing, and 1025 decreasing (Figure A1.S3).

Comparison of the effects of cohesin on transcripts to its binding pattern in BG3 cells argues that many of the effects of Nipped-B and cohesin on gene expression are direct. To ensure that we examined genes that respond consistently, we analyzed transcripts that showed 2-fold or greater increases or decreases in two or more RNAi treatments. By these criteria, 340 transcripts increase, and 414 decrease. 333 of the up-regulated and 407 of the down-regulated genes are euchromatic, allowing us to determine

cohesin and RNA polymerase (PolII) binding from chromatin immunoprecipitation data.

Justified by their genome-wide co-localization [15], we combined the ChIP-chip data for Nipped-B and Smc1 and identified the genes in which these proteins bind within the transcription units at $p \leq 10^{-3}$. By these criteria, 57% (189/333) of the genes that increase, and 36% (146/407) of the genes that decrease in expression bind cohesin (Table A1.S5), which is a significant difference ($p = 9.7 \times 10^{-9}$). PolII binding does not differ, with binding to 68% (225/333) of the increasing and 66% (268/407) of the decreasing genes (Table A1.S5). It is not unexpected that PolII binding is not detected in some cases because many genes are expressed at low levels and have low polymerase density. PolII binding is detected more frequently with the cohesin-binding genes, in 83% of the increasing and 82% of the decreasing genes (Table A1.S5). We conclude that more genes that increase in expression with cohesin RNAi bind cohesin compared to genes that decrease.

Both increasing and decreasing genes bind cohesin at a higher than average frequency. Genome-wide, 19% (816/4282) of PolII-binding genes also bind cohesin, compared to 70% (157/225) of the PolII-binding genes that increase in expression, and 45% (120/268) of the PolII-binding genes that decrease (Table A1.S5). This argues that cohesin directly affects expression, and that negative effects are more common than positive. These data also indicate that many changes in expression that occur with cohesin RNAi are indirect.

Analysis of cohesin-binding genes further argues that the large increases in E(spl)-C and *invected-engrailed* transcripts that occur with cohesin knockdown are unique. Of the 816 genes in BG3 cells that bind both cohesin and PolII, 804 are detected

by the expression microarray. 341 (42%) of these increase in expression by 20% or more with Rad21 knockdown, and 136 (17%) decrease 20% or more (Figure A1.S4). 54 (7%) are not detectably expressed, and 273 (34%) change less than 20% in expression (Figure A1.S4). For genes that increase 20% or more, the median increase is 50%. For the genes that decrease 20% or more, the median decrease is 35%. Thus the effect on expression of most cohesin-binding genes is less than 2-fold.

Cohesin Has Minor Effects on Genes Involved in Translation and Cell Division

The top gene ontology (GO) categories for genes that increase in expression with cohesin RNAi involve development, while the top categories for decreasing transcripts involve protein translation (Figure A1.S3, Table A1.S6). All ribosomal protein transcripts decrease an average of 15%, and all aminoacyl tRNA synthetase transcripts decrease an average of 33% (Table A1.S4). The most significant cell division category is mitotic spindle elongation (Table A1.S6), but most genes in this case encode ribosomal proteins. There are slight increases, all less than 2-fold, in transcripts for *cyclin B*, some cohesion factors and condensin subunits, consistent with a mild G2/M delay [33].

Discussion

Cohesin Regulates the *Enhancer of split* and *invected-engrailed* Gene Complexes in a Cell-Specific Manner

Here we show that in BG3 cells derived from central nervous system, the E(spl)-C, and the complex containing *invected* and *engrailed* share exceptional attributes: (a) cohesin binds over the entire gene complex and not just to individual genes, (b) cohesin binds throughout a large H3K27Me3 domain, and (c) they show unusually large increases in transcription when cohesin is reduced. We posit, therefore, that cohesin directly regulates these gene complexes.

This is supported by the contrasts in histone modification, cohesin binding, and the response to cohesin between BG3 and Sg4 cells. In Sg4 cells, cohesin binds only three of the active E(spl)-C genes, there is no H3K27Me3, and expression not substantially affected by cohesin. Thus the effect of cohesin on the E(spl)-C correlates with presence of cohesin and H3K27Me3 domains. The *invected-engrailed* complex in Sg4 cells shows the typical pattern for PcG silenced genes. It is coated by H3K27Me3, there is no cohesin, and it is silent before or after cohesin RNAi. Thus, we suggest that in BG3 cells, cohesin prevents complete silencing of *invected* and *engrailed* by PcG proteins, and/or that lack of silencing promotes cohesin binding. This latter possibility alone seems unlikely, given that many non-silenced and active genes do not bind cohesin, and that cohesin domains that extend over entire gene complexes are rare. For instance, only selected active E(spl)-C genes bind cohesin in Sg4 cells, in which there is no H3K27Me3, but the entire complex binds cohesin in BG3 cells, when it is also coated by H3K27Me3, indicating that lack of silencing or gene expression by itself is insufficient to establish the cohesin domain. We currently do not know the factors that determine when

and where a cohesin domain is established.

The similarities in chromatin structure and hypersensitivity to cohesin between the E(spl)-C and *invected-engrailed* complexes in BG3 cells lead us to speculate that in cases of cohesin and H3K27Me3 overlap, cohesin helps create an intermediate chromatin structure with aspects of both silenced and active regions (Figure A1.8). Such a dual role is consistent with the biphasic effects of Nipped-B and Rad21 RNAi on E(spl)-C transcription. When cohesin levels are reduced, silencing becomes temporarily stronger, but eventually a specific chromatin structure needed to repress transcription is lost, leading to overexpression. In other regions of cohesin-H3K27Me3 overlap, where genes such as *Psc* and *hth* are expressed at higher levels, the structural balance favors the active state. RNA levels are still increased in these cases by reducing cohesin levels, however, indicating that transcription is still restricted. At present, we do not know if cohesin binding is reduced selectively at specific sites when cohesin or Nipped-B dosage is only slightly reduced, which might contribute to biphasic effects at some genes. The lack of an effect of CP190 insulator protein on E(spl)-C and *invected-engrailed* expression argues against the possibility that changes in insulator activity contribute to the changes in E(spl)-C and *invected-engrailed* transcription that occur with cohesin knockdown.

In *S. cerevisiae*, cohesin inhibits spreading of SIR silencing proteins and establishment of silencing [45,46], suggesting that cohesin might have a similar effect on PcG function at the E(spl)-C and *invected-engrailed* complex. Cohesin binds the silent *HMR* mating type locus [47,48], where it helps form a chromatin boundary [45], and mediate sister cohesion [49,50]. It remains to be determined if cohesin's functions at *HMR* are analogous to its roles at E(spl)-C or *invected-engrailed*, but we note that the

H3K27Me3 mark at *invected-engrailed* extends far beyond the cohesin domain at one end, arguing that cohesin does not form a chromatin boundary.

The finding that H3K27Me3 coats the E(spl)-C and *invected-engrailed* complex in BG3 cells, and that many of the genes in these two complexes bind PolII, raises the question if they are equivalent to bivalent genes in mammals. Including the E(spl)-C and *invected-engrailed*, and the five other genes in regions of cohesin-H3K27Me3 overlap, only 13% of the 480 genes marked by H3K27Me3 in BG3 cells bind PolII, and the vast majority of marked genes are not detectably expressed above background levels. Bivalent genes are defined by the simultaneous presence of the H3K27Me3 mark made by E(z) orthologs at silenced genes, and the histone H3 lysine 4 trimethylation (H3K4Me3) modification made by Trithorax orthologs at active genes [51-53]. Bivalent genes are frequent in embryonic stem cells, but also occur in lineage-restricted cells [53]. Like the E(spl)-C and *invected-engrailed* complex in BG3 cells, many bivalent genes encode transcription factors and are expressed at modest levels [52,54]. The *invected-engrailed* complex in BG3 cells has both H3K4Me3 and H3K27Me3 modifications, but the E(spl)-C shows only a little H3K4Me3 (Y.B. Schwartz, T.G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, V. Pirrotta, submitted). Thus *invected-engrailed* matches the original definition of bivalent genes.

Does Cohesin Regulate the E(spl)-C and *invected-engrailed* In Vivo?

The enhancement of N^{spl-1} mutant phenotypes by *Rad21* (*vtd*) mutations reported here supports the idea that cohesin restricts E(spl)-C transcription during eye and bristle development, because these are the phenotypic changes seen when E(spl)-C activity is increased by gene duplication, forced overexpression, or hypermorphic mutations, and

opposite of what is caused by an increase in Notch signaling or decrease in E(spl)-C dosage [37-41].

Heterozygous *Nipped-B* mutations suppress the N^{spl-1} eye phenotype, suggesting that they either reduce E(spl)-C expression or increase Notch signaling. Because heterozygous *Nipped-B* null mutations only reduce *Nipped-B* mRNA by 25% [8], this is consistent with an in vivo biphasic effect on E(spl)-C transcription similar to that seen in BG3 cells. Based on the genome-wide analysis in BG3 cells, which shows that Nipped-B and cohesin regulate the same genes to similar extents, it is unlikely that Nipped-B and Rad21 have opposing effects on eye development by regulating different genes. Also, *Nipped-B* mutations do not affect other sensitive *Notch* mutant phenotypes, arguing that the effect on N^{spl-1} is not through increasing Notch signaling [9]. Given the essential nature of cohesin in cell division, and the complex spatial and temporal pattern of E(spl)-C expression in vivo, it will not be simple to confirm that Nipped-B and cohesin directly affect the levels of specific E(spl)-C transcripts in vivo, or rule out potential indirect effects. Indeed, given the contrast in binding of cohesin to the E(spl)-C between BG3 and Sg4 cells, in vivo effects of cohesin likely occur in only a select population of E(spl)-C expressing cells.

For similar reasons, it will also not be straightforward to confirm that PcG proteins regulate the E(spl)-C in vivo. Effects of PcG on E(spl)-C function have not been reported, and genome-wide mapping in other cell lines, whole organisms, or imaginal discs has not revealed that the E(spl)-C gene is a PcG target [30,31,55-57]. Nonetheless, the H3K27Me3 pattern and the effects of Pc knockdown on E(spl)-C expression in BG3 cells argue strongly that E(spl)-C is a PcG target, although this may occur only in a small

fraction of cells in vivo.

It is unknown if *invected* and *engrailed* are regulated by cohesin in vivo. Our results suggest that this may occur in cells in which *engrailed* is active, but partially repressed by PcG proteins, such as the posterior compartment of the wing imaginal disc [58]. No dominant effects of *Nipped-B* or cohesin mutations on compartment formation have been observed in otherwise wild-type flies, but the feedback loop at the wing anterior-posterior boundary that controls *engrailed, hedgehog, patched, wingless* and *decapentaplegic* expression [59] may prevent or counteract increases in *engrailed* expression. The feedback mechanisms may be unbalanced in $hedgehog^{Moonrat}$ mutants, in which ectopic *hedgehog* expression in the anterior compartment causes overgrowth [60,61]. *Rad21* (*vtd*) and *Nipped-B* mutations dominantly suppress this overgrowth [12,62], and one possibility is that increased *engrailed* expression helps restore the autoregulatory loop.

Do Genes Hypersensitive to Cohesin Contribute to Cornelia de Lange Syndrome (CdLS)?

Heterozygous loss-of-function mutations in the *Nipped-B-Like* (*NIPBL*) ortholog of *Nipped-B* cause CdLS, characterized by slow growth, mental retardation, autistic features, craniofacial abnormalities, and structural defects in limbs, gut, heart and kidney [63,64]. Mutations that change amino acid residues in the Smc1 or Smc3 cohesin subunits cause milder CdLS [65,66]. Cells from CdLS individuals do not have significant defects in chromatid cohesion [67-69], and *NIPBL* mRNA is only reduced by 15 to 30% in cells from CdLS individuals [70,71], indicating that the developmental deficits arise from changes in gene expression.

Relative to healthy controls, over a thousand genes that are differentially expressed in CdLS lymphocyte cell lines with *NIPBL* mutations or mutant Smc1 [71]. As with cohesin knockdown in Drosophila BG3 cells, some genes increase in expression and some decrease. Most changes in lymphocytes, however, are less than 2-fold, and the largest effect is less than 4-fold. It is unknown if lymphocytes contain significant overlaps of cohesin and H3K27Me3, and therefore whether or not they might have hypersensitive genes similar to those in BG3 cells. Given the small reductions in cohesion factor activity that cause CdLS, the findings in BG3 cells suggest that genes that are hypersensitive to cohesin in only a subset of cells are the most likely to be strongly affected, and significantly alter development.

Materials and Methods

Cell Culture and RNAi

BG3 cells were cultured in Schneider's media with 10% FCS and 10 μg per ml insulin. Sg4 cells were grown in Schneider's containing 10% FCS. For RNAi, cells were plated at $5x10^6$ cells per 3 cm well for BG3 cells, and $3x10^6$ for Sg4 cells. Media was replaced with 1 ml of Express Five SFM (Invitrogen) with 1% FCS, (and 10 μg per ml insulin for BG3 cells). For cohesion factors and Polycomb, from 0.7 to 40 μg of dsRNA was added per well, and 80 μg was used for CP190 knockdown. Media was adjusted to 3 ml and 10% FCS with Schneider's media after 2 hrs. Cells were replated as needed. Templates for dsRNA synthesis were made by PCR from cDNA or genomic DNA templates using primers with T7 promoters (Table A1.S7). In most experiments, equal amounts of two dsRNAs against each target were used. Both individual dsRNAs knocked down the targets, but knockdown was generally more efficient with a mixture. All dsRNA sequences were scanned against the genome to avoid off-target effects. To determine transcript half-lives, actinomycin D was added to cultures at 5 μg per ml, RNA was extracted every 30 min up to 2 hours, and half-lives were calculated assuming exponential decay.

RNA Quantification

Total RNA was isolated using Trizol (Invitrogen), treated with DNase I (Epicentre), chloroform extracted, ethanol precipitated and dissolved in water. cDNA was synthesized using random hexamer primers and SuperScript VILO reverse transcriptase

(Invitrogen). Transcripts were quantified using Sybr green real-time PCR (Clontech) and gene-specific primers (Table A1.S8) calibrated with genomic DNA. RNA levels were calculated adjusting for amplification efficiency [72] and normalizing to internal *RpL32* transcripts and external genomic DNA standards. Standard errors of the mean were calculated using all PCR replicates from all biological replicates.

Protein Extracts and Western Blots

Cells were washed in PBS, lysed in RIPA buffer (5 µl per 10^6 cells), insoluble material removed by centrifugation, and extracts were stored at -80°. Nipped-B, Smc1, SA, Rad21, Polycomb, and CP190 proteins were quantified by SDS-PAGE western blots using chemiluminescence imaging with Actin as a standard and previously described antisera [6,7,15,20,73].

Metaphase Spreads

Cells ($3x10^6$) were incubated in media with 3 mg per ml colchicines for 4 hr, washed in phosphate-buffered saline (PBS), suspended in hypotonic (1% sodium citrate) for 4 min, collected by centrifugation, suspended in 0.1 ml hypotonic and fixed with 1 ml ice-cold methanol:acetic acid (3:1). Fixed cells were suspended in 60 µl of methanol:acetic acid, dropped onto a microscope slide from a distance of 50 to 60 cm, and covered with a coverglass. Slides were frozen on dry ice for 20 min, and rinsed with PBST (PBS with 1% Triton X-100) 3 times after removing the coverslip. Chromosomes were stained with 0.5 µg DAPI per ml in PBS for 10 min, rinsed with PBST, mounted in BioRad FluoroGard, and observed by fluorescence microscopy.

Effects of *Nipped-B* and *Rad21* (*vtd*) Mutations on N^{spl-1} Mutant Phenotypes

w^a N^{spl-1} females were crossed to wild-type males or males with *Nipped-B* and *vtd* mutations over balancers with dominant markers at 25°. The anterior-posterior diameter of the eyes of male progeny were measured with a reticule in a dissection microscope, and scutellar macrochaete were counted.

Genome-Wide Transcript Analysis

Five μg of total RNA purified by Qiagen RNeasy minicolumns was used to make cRNA probes using Affymetrix GeneChip HT One-Cycle Target Labeling and Controls Kit according to the manufacturer's instructions. Probes were hybridized to Affymetrix GeneChip *Drosophila* Genome 2.0 arrays, processed and scanned using Affymetrix procedures. Quality metrics for each array were monitored by spike-in labeling controls and hybridization/staining controls using Microarray Suite 5.0 (MAS5) algorithms from GeneChip® Operating Software v1.4, (GCOS) (Affymetrix, Inc). Probe cell intensities for each array were normalized using GCRMA algorithms, which consist of background adjustment and quantile normalization, accounting for probe GC content [74]. Normalization was executed using the R statistical environment [R Foundation for Statistical Computing, Vienna, 2007; ISBN 3-900051-07-0; www.R-project.org] and the Bioconductor package (www.bioconductor.org) [75]. Transcript levels from Rad21 and Nipped-B RNAi treatments were compared to those of mock RNAi controls at 3 and >3 days (4 and 6 days) (N=4 per RNAi comparison; N=2 per treatment condition). A balanced 2-way ANOVA was performed on GCRMA-normalized $\log_2$ signal intensities to assess expression variability with regard to RNAi treatment (FDR ≤ 0.1) [76,77]. Differentially expressed groups were analyzed for gene ontology enrichment using

Fisher's exact test in the GOEAST package [78]. The data are available in the GEO database (accession no. GSE16152).

Correlation of Chromatin Immunoprecipitation and Gene Expression Data

The Nipped-B, Smc1, RNA polymerase II, H3K27Me3 and control cel files for BG3 and Sg4 cell chromatin immunoprecipitations (GEO acc. no. GSE9248; ArrayExpress acc. no. E-MEXP-535) were processed using MAT [79] to generate cohesin-Nipped-B, H3K27Me3, and PolII bed files at $p \leq 10^{-3}$ that were visualized using the Affymetrix Integrated Genome Browser. Transcription units that overlap cohesin-Nipped-B, H3K27Me3, and PolII binding regions were identified using the April 2006 genome annotations [80].

Acknowledgments

We thank Rick Jones for reagents and helpful discussions, Marek Bartkuhn and Rainer Renkawitz for reagents, Judy Kassis and Jim Jaynes for helpful discussions, and Gary Stormo and Michael Rauchman for comments on the manuscript.

Financial Disclosure

This work was supported by grants from the NIH (R01 GM055683, P01 HD052683, www.nih.gov) and the March of Dimes (FY05-103, www.marchofdimes.com) to D.D. G.S. was supported by NIH grants R01 HG00249 and T32 GM008802 (Gary Stormo,

PI). The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Guacci V, Koshland D, Strunnikov A (1997) A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in *S. cerevisiae*. Cell 91: 47-57.
2. Michaelis C, Ciosk R, Nasmyth K (1997) Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. Cell 91: 35-45.
3. Peters JM, Tedeschi A, Schmitz J (2008) The cohesin complex and its roles in chromosome biology. Genes Dev 22: 3089-3114.
4. Haering CH, Farcas AM, Arumugam P, Metson J, Nasmyth K (2008) The cohesin ring concatenates sister DNA molecules. Nature 454: 297-301.
5. Ivanov D, Nasmyth K (2005) A topological interaction between cohesin rings and a circular minichromosome. Cell 122: 849-860.
6. Dorsett D, Eissenberg JC, Misulovin Z, Martens A, Redding B, McKim K (2005) Effects of sister chromatid cohesion proteins on *cut* gene expression during wing development in Drosophila. Development 132: 4743-4753.
7. Gause M, Webber HA, Misulovin Z, Haller G, Rollins RA, et al. (2008) Functional links between Drosophila Nipped-B and cohesin in somatic and meiotic cells. Chromosoma 117: 51-66.
8. Rollins RA, Korom M, Aulner N, Martens A, Dorsett D (2004) Drosophila

Nipped-B protein supports sister chromatid cohesion and opposes the stromalin/Scc3 cohesion factor to facilitate long-range activation of the *cut* gene. Mol Cell Biol 24: 3100-3111.

9. Rollins RA, Morcillo P, Dorsett D (1999) Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the *cut* and *Ultrabithorax* genes. Genetics 152: 577-593.

10. Pauli A, Althoff F, Oliveira RA, Heidmann S, Schuldiner O, et al. (2008) Cell-type-specific TEV protease cleavage reveals cohesin functions in Drosophila neurons. Dev Cell 14: 239-251.

11. Schuldiner O, Berdnik D, Levy JM, Wu JS, Luginbuhl D, et al. (2008) piggyBac-based mosaic screen identifies a postmitotic function for cohesin in regulating developmental axon pruning. Dev Cell 14: 227-238.

12. Hallson G, Syrzycka M, Beck SA, Kennison JA, Dorsett D, et al. (2008) The Drosophila cohesin subunit Rad21 is a trithorax group (trxG) protein. Proc Natl Acad Sci U S A 105: 12405-12410.

13. Kennison JA, Tamkun JW (1988) Dosage-dependent modifiers of *Polycomb* and *Antennapedia* mutations in Drosophila. Proc Natl Acad Sci U S A 85: 8136-8140.

14. Horsfield JA, Anagnostou SH, Hu JK, Cho KH, Geisler R, et al. (2007) Cohesin-dependent regulation of Runx genes. Development 134: 2639-2649.

15. Misulovin Z, Schwartz YB, Li XY, Kahn TG, Gause M, et al. (2008) Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome. Chromosoma 117: 89-102.

16. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, et al. (2008) Cohesin mediates

transcriptional insulation by CCCTC-binding factor. Nature 451: 796-801.

17. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, et al. (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 132: 422-433.

18. Mishiro T, Ishihara K, Hino S, Tsutsumi S, Aburatani H, et al. (2009) Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J Mar 26 [Epub ahead of print]

19. Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, et al. (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature May 20 [Epub ahead of print]

20. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, et al. (2007) CTCF genomic binding sites in *Drosophila* and the organisation of the bithorax complex. PLoS Genet 3: e112.

21. Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, et al. (2009) Active promoters and insulators are marked by the centrosomal protein 190. EMBO J Feb 19 [Epub ahead of print]

22. Ui K, Nishihara S, Sakuma M, Togashi S, Ueda R, et al. (1994) Newly established cell lines from Drosophila larval CNS express neural specific characteristics. In Vitro Cell Dev Biol Anim 30A: 209–216.

23. Bray SJ (1997) Expression and function of Enhancer of split bHLH proteins during Drosophila neurogenesis. Perspect Dev Neurobiol 4: 313-323.

24. Brower DL (1986) *engrailed* gene expression in Drosophila imaginal discs. EMBO J 5: 2649-2656.

25. Coleman KG, Poole SJ, Weir MP, Soeller WC, Kornberg T (1987) The *invected* gene of *Drosophila*: sequence analysis and expression studies reveal a close kinship to the *engrailed* gene. Genes Dev 1: 19-28.

26. Kornberg T, Sidén I, O'Farrell P, Simon M (1985) The *engrailed* locus of Drosophila: in situ localization of transcripts reveals compartment-specific expression. Cell 40: 45-53.

27. DeVido SK, Kwon D, Brown JL, Kassis JA (2008) The role of Polycomb-group response elements in regulation of *engrailed* transcription in Drosophila. Development 135: 669-676.

28. Drees B, Ali Z, Soeller WC, Coleman KG, Poole SJ, Kornberg T (1987) The transcription unit of the Drosophila *engrailed* locus: an unusually small portion of a 70,000 bp gene. EMBO J 6: 2803-2809.

29. Fujioka M, Yusibova GL, Zhou J, Jaynes JB (2008) The DNA-binding Polycomb-group protein Pleiohomeotic maintains both active and repressed transcriptional states through a single site. Development 135: 4131-4139.

30. Kwong C, Adryan B, Bell I, Meadows L, Russell S, et al. (2008) Stability and dynamics of polycomb target sites in *Drosophila* development. PLoS Genet 4: e1000178.

31. Oktaba K, Gutiérrez L, Gagneur J, Girardot C, Sengupta AK, et al. (2008) Dynamic regulation by polycomb group protein complexes controls pattern formation and the cell cycle in Drosophila. Dev Cell 15: 877-889.

32. Papp B, Müller J (2006) Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. Genes Dev 20:

2041-2054.

33. Vass S, Cotterill S, Valdeolmillos AM, Barbero JL, Lin E, et al. (2003) Depletion of Drad21/Scc1 in Drosophila cells leads to instability of the cohesin complex and disruption of mitotic progression. Curr Biol 13: 208-218.

34. Lai EC, Tam B, Rubin GM (2005) Pervasive regulation of *Drosophila* Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. Genes Dev 19: 1067-1080.

35. Moazed D, O'Farrell PH (1992) Maintenance of the *engrailed* expression pattern by *Polycomb* group genes in *Drosophila*. Development 116: 805-810.

36. Bushey AM, Ramos E, Corces VG (2009) Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev 23: 1338-1350.

37. Nagel AC, Preiss A (1999) *Notch*^{spl} is deficient for inductive processes in the eye, and *E(spl)*^{D} enhances *split* by interfering with proneural activity. Dev Biol 208: 406-415.

38. Knust E, Bremer KA, Vässin H, Ziemer A, Tepass U, Campos-Ortega JA (1987) The *enhancer of split* locus and neurogenesis in *Drosophila melanogaster*. Dev Biol 122: 262-273.

39. Ligoxygakis P, Yu SY, Delidakis C, Baker NE (1998) A subset of notch functions during Drosophila eye development require Su(H) and the *E(spl)* gene complex. Development 125: 2893-2900.

40. Welshons WJ (1956) Dosage experiments with *split* mutants in the presence of an enhancer of *split*. D I S 30: 157-158.

41. Shepard SB, Broverman SA, Muskavitch MA (1989) A tripartite interaction among alleles of *Notch*, *Delta*, and *Enhancer of split* during imaginal development of *Drosophila melanogaster*. Genetics 122: 429-438.

42. Cooper MT, Tyler DM, Furriols M, Chalkiadaki A, Delidakis C, Bray S (2000) Spatially restricted factors cooperate with Notch in the regulation of *Enhancer of split* genes. Dev Biol 221: 390-403.

43. Nellesen DT, Lai EC, Posakony JW (1999) Discrete enhancer elements mediate selective responsiveness of *enhancer of split* complex genes to common transcriptional activators. Dev Biol 213: 33-53.

44. Maeder ML, Polansky BJ, Robson BE, Eastman DA (2007) Phylogenetic footprinting analysis in the upstream regulatory regions of the Drosophila *Enhancer of split* genes. Genetics 177: 1377-1394.

45. Donze D, Adams CR, Rine J, Kamakaka RT (1999) The boundaries of the silenced *HMR* domain in *Saccharomyces cerevisiae*. Genes Dev 13: 698-708.

46. Lau A, Blitzblau H, Bell SP (2002) Cell-cycle control of the establishment of mating-type silencing in *S. cerevisiae*. Genes Dev 16: 2935-2945.

47. Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, et al. (2007) Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae*. PLoS Biol 2: e259.

48. Lengronne A, Katou Y, Mori S, Yokobayashi S, Kelly GP, et al. (2004) Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature 430: 573-578.

49. Chang CR, Wu CS, Hom Y, Gartenberg MR (2005) Targeting of cohesin by

transcriptionally silent chromatin. Genes Dev 19: 3031-3042.

50. Dubey RN, Gartenberg MR (2007) A tDNA establishes cohesion of a neighboring silent chromatin domain. Genes Dev 21: 2150-2160.

51. Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, et al. (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532-538.

52. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

53. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

54. Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, et al. (2007) Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 9: 1428-1435.

55. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, et al. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7: e13.

56. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. Nat Genet 38: 700-705.

57. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, et al. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in *Drosophila melanogaster*. Nat Genet 38: 694-699.

58. Randsholt NB, Maschat F, Santamaria P (2000) *polyhomeotic* controls *engrailed*

expression and the hedgehog signaling pathway in imaginal discs. Mech Dev 95: 89-99.

59. Sanicola M, Sekelsky J, Elson S, Gelbart WM (1995) Drawing a stripe in Drosophila imaginal disks: negative regulation of *decapentaplegic* and *patched* expression by *engrailed*. Genetics 139: 745-756.

60. Felsenfeld AL, Kennison JA (1995) Positional signaling by hedgehog in *Drosophila* imaginal disc development. Development 121: 1-10.

61. Tabata T, Kornberg TB (1994) Hedgehog is a signaling protein with a key role in patterning Drosophila imaginal discs. Cell 76: 89-102.

62. Schulze S, Sinclair DA, Silva E, Fitzpatrick KA, Singh M, et al. (2001) Essential genes in proximal 3L heterochromatin of *Drosophila melanogaster*. Molec Gen Genet 264: 782-789.

63. Krantz ID, McCallum J, DeScipio C, Kaur M, Gillis LA, et al. (2004) Cornelia de Lange syndrome is caused by mutations in *NIPBL*, the human homolog of *Drosophila melanogaster Nipped-B*. Nat Genet 36: 631-635.

64. Tonkin ET, Wang TJ, Lisgo S, Bamshad MJ, Strachan T (2004) *NIPBL*, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. Nat Genet 36: 636-641.

65. Deardorff MA, Kaur M, Yaeger D, Rampuria A, Korolev S, et al. (2007) Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation. Am J Hum Genet. 80: 485-494.

66. Musio A, Selicorni A, Focarelli ML, Gervasini C, Milani D, et al. (2006) X-linked

Cornelia de Lange syndrome owing to *SMC1L1* mutations. Nat Genet 38: 528-530.

67. Kaur M, DeScipio C, McCallum J, Yaeger D, Devoto M, et al. (2005) Precocious sister chromatid separation (PSCS) in Cornelia de Lange syndrome. Am J Med Genet A 138: 27-31.

68. Revenkova E, Focarelli ML, Susani L, Paulis M, Bassi MT, et al. (2009) Cornelia de Lange syndrome mutations in SMC1A or SMC3 affect binding to DNA. Hum Mol Genet 18: 418-427.

69. Vrouwe MG, Elghalbzouri-Maghrani E, Meijers M, Schouten P, Godthelp BC, et al. (2007) Increased DNA damage sensitivity of Cornelia de Lange syndrome cells: evidence for impaired recombinational repair. Hum Mol Genet 16: 1478-1487.

70. Borck G, Zarhrate M, Cluzeau C, Bal E, Bonnefont JP, et al. (2006) Father-to-daughter transmission of Cornelia de Lange syndrome caused by a mutation in the 5' untranslated region of the *NIPBL* gene. Hum Mutat 27: 731-735.

71. Liu J, Zhang Z, Bando M, Itoh T, Deardorff MA, et al. (2009) Transcriptional dysregulation in NIPBL and cohesin mutant human cells. PLoS Biol 7: e1000119.

72. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29: e45.

73. Wang L, Brown JL, Cao R, Zhang Y, Kassis JA, Jones RS (2004) Hierarchical recruitment of polycomb group silencing complexes. Mol Cell 14: 637-646.

74. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. J American

Statistical Association 99: 909-917.

75. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

76. Pavlidis P (2003) Using ANOVA for gene selection from microarray studies of the nervous system. Methods 31: 282-289.

77. Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

78. Zheng Q, Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36: W358-W363.

79. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, et al. (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A 103: 12457-12462.

80. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, et al. (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol. 3: RESEARCH0079.

Figure Legends

Figure A1.1. *Enhancer of split* and *invected-engrailed* gene complexes.

The *Enhancer of split* complex [E(spl)-C] (top) contains twelve genes (blue): *HLHmδ*, *HLHmγ*, *HLHmβ*, *mα*, *m1*, *m2*, *HLHm3*, *m4*, *HLHm5*, *m6*, *HLHm7*, and *E(spl)m8*. Nucleotide numbering is from the April 2006 genome (Berkeley Drosophila Genome Project). Genes above the scale are transcribed from left to right, and those below from right to left. Tracks above the gene diagrams show chromatin immunoprecipitation data for histone H3 lysine 27 trimethylation (H3K27Me3), RNA polymerase II (PolII) and combined cohesin and Nipped-B binding (cohesin-Nipped-B) for Sg4 (red) and BG3 cells (black) [15,56, Y.B. Schwartz, T.G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta, submitted). Bars below each track show regions that bind at $p \leq 10^{-3}$, as determined using the MAT program. The bottom shows the same for the *invected-engrailed* complex.

Figure A1.2. Regulation of the E(spl)-C and *invected-engrailed* complex by cohesin and Nipped-B.

(A) Transcripts for the E(spl)-C and *invected-engrailed* complex in BG3 (black) and Sg4 (red) cells quantified by RT-PCR and normalized to *RpL32*. The *HLHmδ* level in BG3 cells is defined as 1 unit, and all transcripts are normalized to this value. By comparison to genomic DNA standards, *HLHmδ* transcripts in BG3 cells are 8,400-fold less than *RpL32* transcripts. BG3 values are the average of three RNA preparations, and Sg4 values

are the average of two. Standard errors were calculated using all RT-PCR replicates from all biological replicates.

(B) Rad21 RNAi time course, for Rad21 protein (blue diamonds, 100% starting), and fold-increases for the *HLHmδ* (red squares) and *invected* (green triangles) transcripts. Similar time courses are seen for *engrailed* and other E(spl)-C transcripts (not shown). Nipped-B knockdown shows similar time courses in Nipped-B protein and E(spl)-C and *invected-engrailed* transcripts (not shown), except that some E(spl)-C transcripts decrease on day 3 (Figure A1.3).

(C) The left panel shows transcript levels in a typical experiment with mock RNAi-treated BG3 cells (black) and BG3 cells six days after Rad21 (blue) or Nipped-B (orange) RNAi treatment. The right panel shows transcript levels in another experiment with mock-treated BG3 cells (black), and BG3 cells treated with Rad21 (blue) or SA (purple) RNAi six days after treatment.

(D) E(spl)-C and *invected-engrailed* transcript levels in mock-RNAi treated Sg4 cells (red), or Sg4 cells after two successive 3 day Rad21 (blue) or Nipped-B (orange) RNAi treatments.

(E) Western blots of whole cell extracts after RNAi treatment. The three left panels show the same blot of BG3 extracts six days after RNAi probed with Nipped-B, Rad21 and Actin antisera. RNAi treatments are indicated at the tops of the lanes. The middle three panels show a blot of Sg4 extracts after two successive 3 day RNAi treatments. The right panels show a blot of BG3 extracts probed with SA, Rad21 and Actin antibodies six days after RNAi.

Figure A1.3. Biphasic changes in E(spl)-C transcripts after Nipped-B and Rad21 knockdown in BG3 cells.

The top panel shows E(spl)-C transcript levels in mock-treated (black) or Nipped-B RNAi treated (orange) BG3 cells three days after treatment. Similar results were obtained in all Nipped-B RNAi experiments. All levels are relative to *HLHmδ* in mock-treated cells. The data shown is an average of two RNAi experiments. The bottom panel shows the indicated E(spl)-C transcript levels three days after treatments with increasing amounts of Rad21 dsRNA that cause different extents of knockdown (mock, 0%; 0.7 μg per 3 cm well, 32%; 1.7 μg, 55%; 3.3 μg, 71%; 6.7 μg, 81%).

Figure A1.4. Effects of Polycomb on E(spl)-C and *invected-engrailed* transcripts in BG3 cells.

The graph shows transcript levels in mock-treated BG3 cells (black) and in Polycomb RNAi-treated cells (gray) six days after treatment. The western blot shows the Polycomb protein knockdown (~70%) on day 6. All transcripts are relative to *HLHmδ* in mock control cells. Similar results were obtained in three experiments.

Figure A1.5. Effects of the CP190 insulator protein on E(spl)-C and *invected-engrailed* transcripts in BG3 cells.

The graph shows transcript levels in mock-treated BG3 cells (black), Rad21 (blue) and CP190 (green) RNAi-treated BG3 cells six days after treatment. The western blot shows the knockdown of CP190 protein on days 4 and 6 (~90%). The unlabeled protein under 72 kD in size that is unaffected by RNAi is a cross-reacting cytoplasmic protein (Marek

Bartkuhn and Rainer Renkawitz, personal communication). Similar results were obtained with 4 and 8 days after CP190 RNAi.

Figure A1.6. Dominant effects of *Nipped-B* and *Rad21* mutations on *Notch-split* (N^{spl-1}) mutant phenotypes.

The top panel compares the eye phenotype in two wild-type backgrounds (wt a, Oregon R; wt b, Canton S), to flies heterozygous for $Nipped\text{-}B^{407}$, $Rad21^{36}$ (vtd^{36}), and $Rad21^{\gamma 26\text{-}6}$ ($vtd^{\gamma 26\text{-}6}$). Eye diameter was measured as shown in the upper right. At least 30 eyes were scored for each genotype. Error bars are standard errors. The bottom panel shows the effects of the heterozygous *Nipped-B* and *Rad21* mutations on the four scutellar macrochaete (large bristles). The number of flies scored for bristles is given above the bars.

Figure A1.7. Genome-wide effects of Rad21 and Nipped-B RNAi on RNA transcripts in BG3 cells.

The top graph shows the effects of Rad21 knockdown on transcript levels ($\log_2$ Rad21/Mock) versus the effects of Nipped-B knockdown ($\log_2$ Nipped-B/Mock), 6 days after RNAi for all 18,770 probes on the microarray. E(spl)-C and *invected-engrailed* transcripts are red. The bottom is an aligned histogram of the effects of Rad21 RNAi, with transcripts that increase 2-fold or more in expression in red, and transcripts that decrease 2-fold or more in green.

Figure A1.8. Speculative model for regulation of gene complexes by cohesin.

The top depicts a PcG-silenced complex contained in a loop created by PRE-PRE interactions. There is little or no transcription, and we posit that the silenced chromatin diameter prevents encirclement by cohesin. The nucleosomes have trimethylation of histone H3 on lysine 27 (green). The middle diagram depicts a gene complex in which cohesin, trithorax group (trxG), transcriptional activators, and PcG proteins combine to create an intermediate chromatin structure with aspects of both silenced and active regions that permits modest transcription (angled arrows); nucleosomes near the transcription start sites also have trimethylation of histone H3 on lysine 4 (pink). Based on the biphasic effects of Nipped-B and cohesin knockdown on some E(spl)-C transcripts, we posit that when cohesin levels are reduced, the chromatin structure first becomes closer to the silenced state, decreasing transcription, and that the higher order structure associated with silencing is eventually lost, leading to unrestrained transcription (bottom).

Figure A1.1

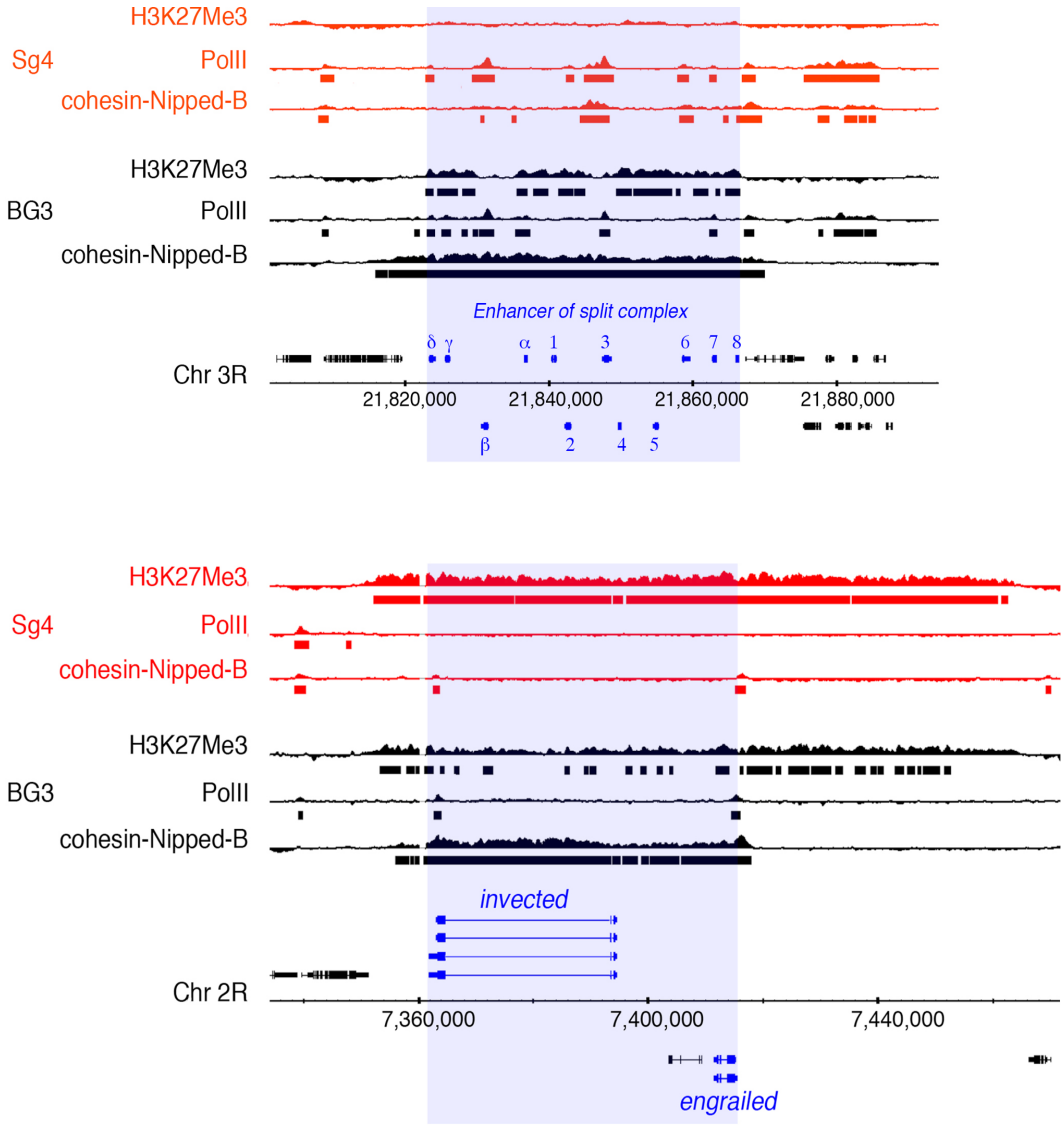

H3K27Me3
Sg4
PolII
cohesin-Nipped-B
H3K27Me3
BG3
PolII
cohesin-Nipped-B
Enhancer of split complex
δ γ α 1 3 6 7 8
β 2 4 5
Chr 3R
21,820,000
21,840,000
21,860,000
21,880,000
H3K27Me3
Sg4
PolII
cohesin-Nipped-B
H3K27Me3
BG3
PolII
cohesin-Nipped-B
invected
Chr 2R
7,360,000
7,400,000
7,440,000
engrailed

Figure A1.2

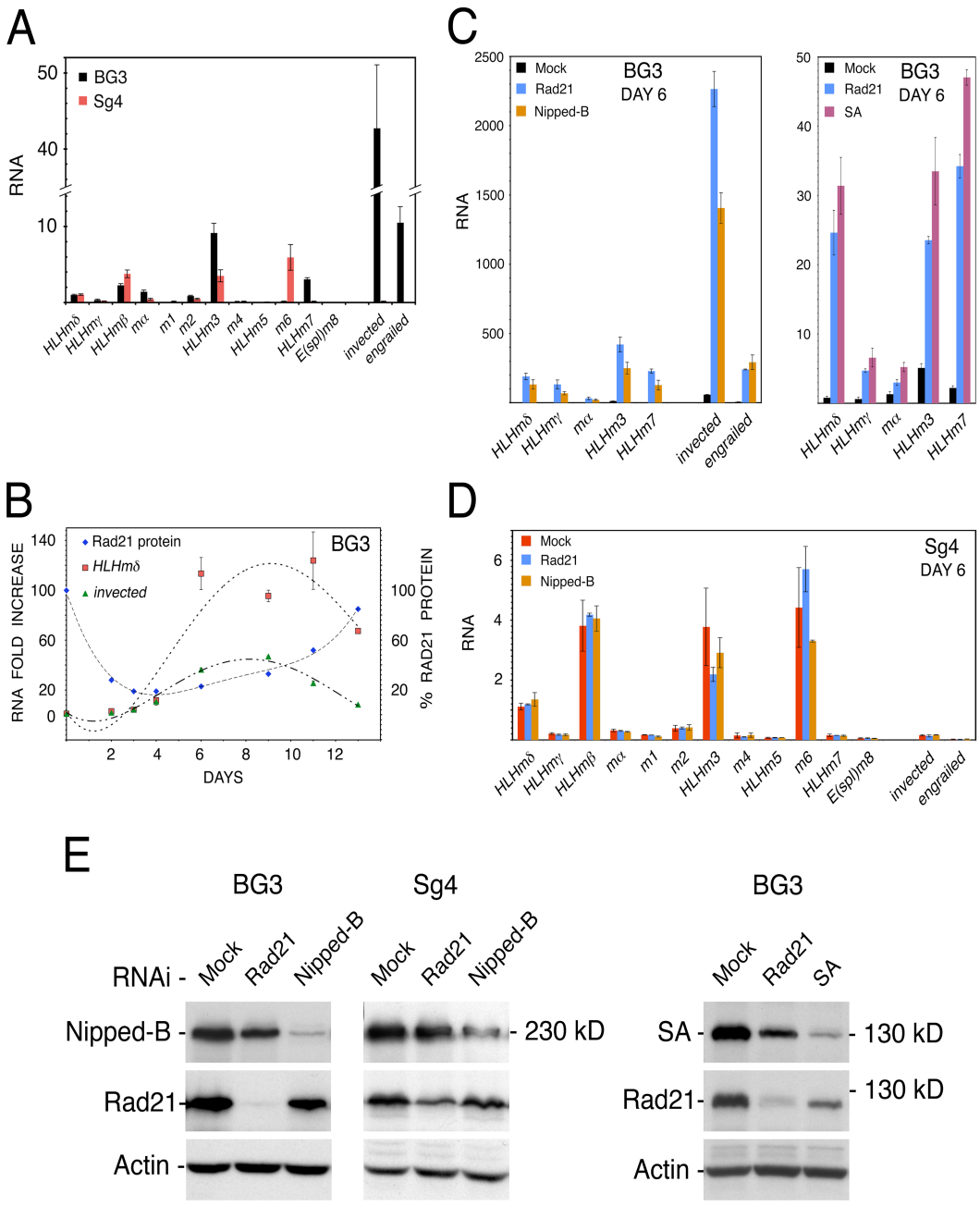

A
BG3
Sg4
RNA
HLHmδ
HLHmγ
HLHmβ
mα
m1
m2
HLHm3
m4
HLHm5
m6
HLHm7
E(spl)m8
invected
engrailed
C
Mock
Rad21
Nipped-B
BG3
DAY 6
RNA
SA
B
Rad21 protein
HLHmδ
invected
BG3
RNA FOLD INCREASE
% RAD21 PROTEIN
DAYS
D
Sg4
DAY 6
E
BG3
Sg4
RNAi -
Mock
Rad21
Nipped-B
SA
Nipped-B -
Rad21-
Actin -
230 kD
130 kD

Figure A1.3

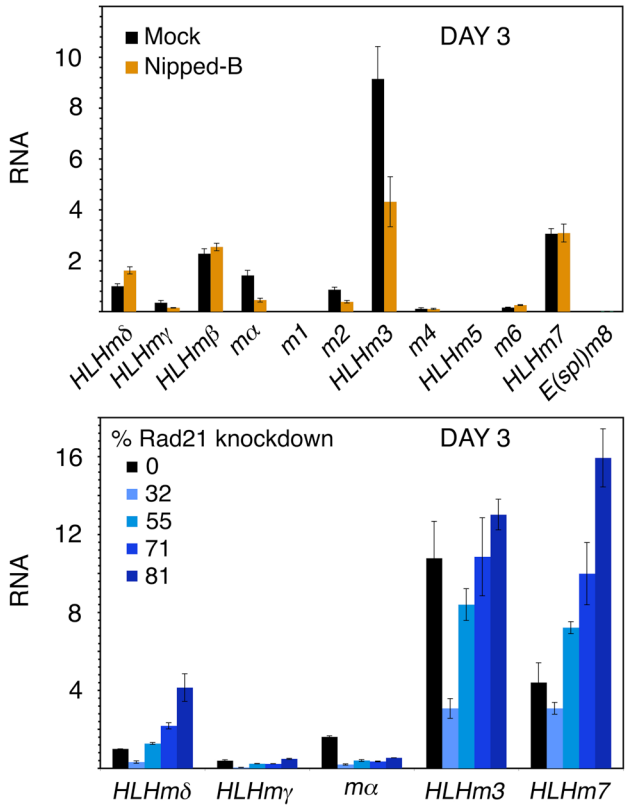

Mock
Nipped-B
DAY 3
RNA
10
8
6
4
2
HLHmδ
HLHmγ
HLHmβ
mα
m1
m2
HLHm3
m4
HLHm5
m6
HLHm7
E(spl)m8
% Rad21 knockdown
0
32
55
71
81
DAY 3
RNA
16
12
8
4
HLHmδ
HLHmγ
mα
HLHm3
HLHm7

Figure A1.4

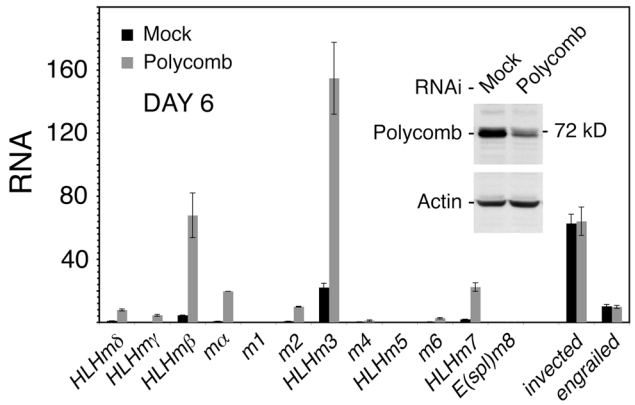

Mock
Polycomb
DAY 6
RNA
160
120
80
40
RNAi - Mock
Polycomb
Polycomb -
- 72 kD
Actin -
HLHmδ
HLHmγ
HLHmβ
mα
m1
m2
HLHm3
m4
HLHm5
m6
HLHm7
E(spl)m8
invected
engrailed

Figure A1.5

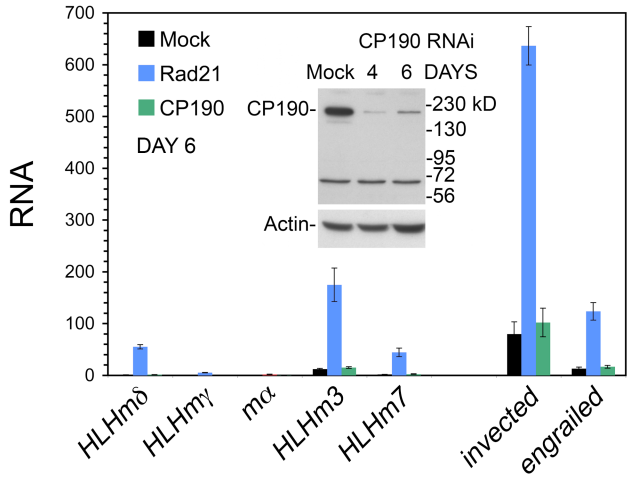

RNA
700
600
500
400
300
200
100
0
Mock
Rad21
CP190
DAY 6
CP190 RNAi
Mock 4 6 DAYS
CP190-
-230 kD
-130
-95
-72
-56
Actin-
HLHmδ
HLHmγ
mα
HLHm3
HLHm7
invected
engrailed

Figure A1.6

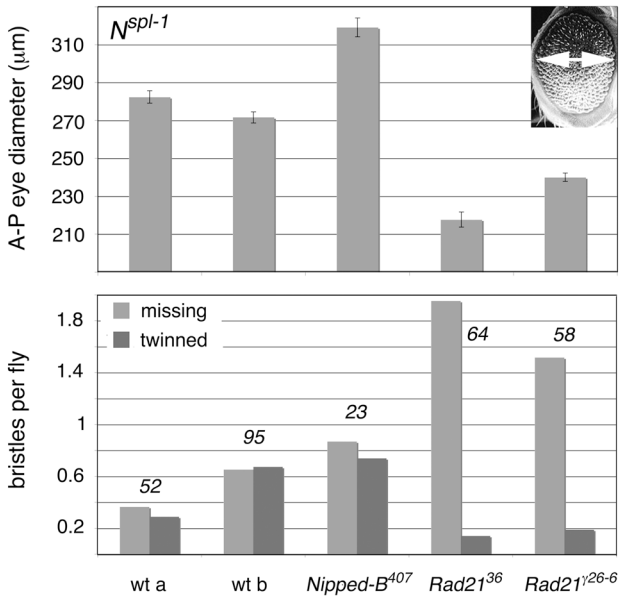

N^spl-1
A-P eye diameter (μm)
310
290
270
250
230
210
bristles per fly
missing
twinned
1.8
1.4
1
0.6
0.2
52
95
23
64
58
wt a
wt b
Nipped-B^407
Rad21^36
Rad21^γ26-6

Figure A1.7

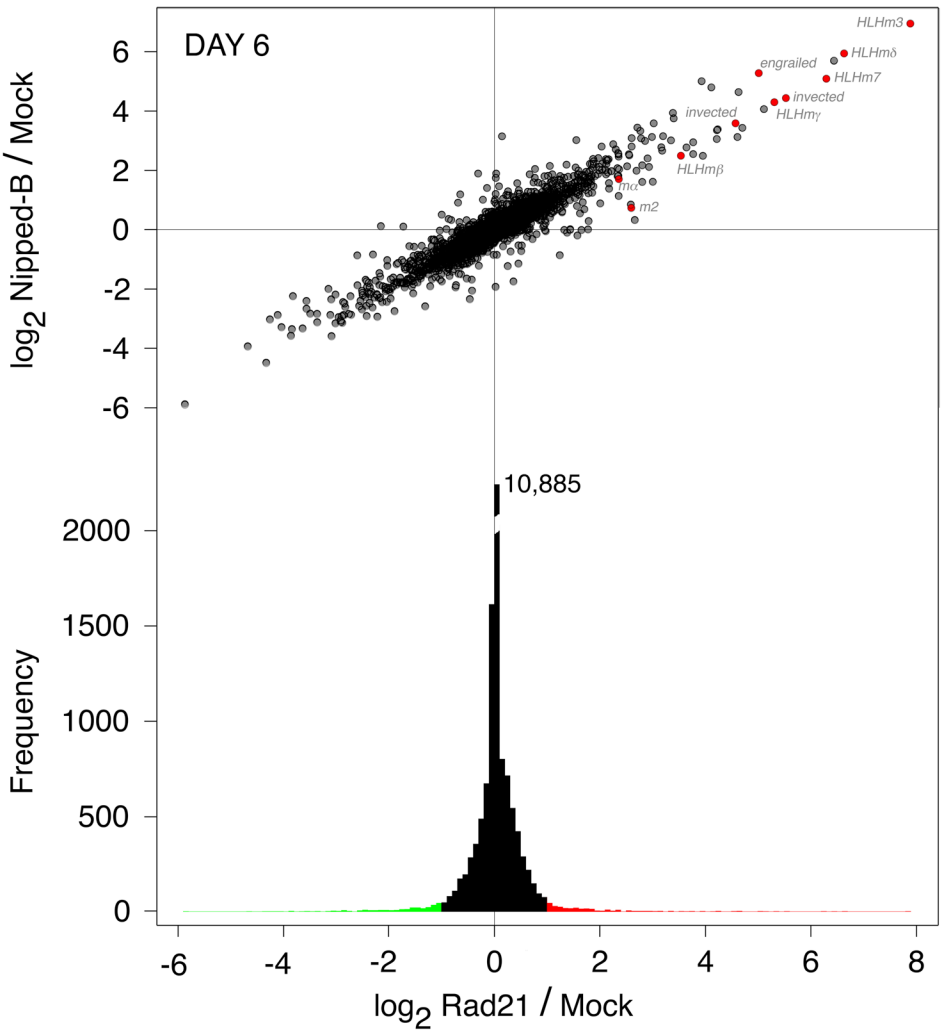
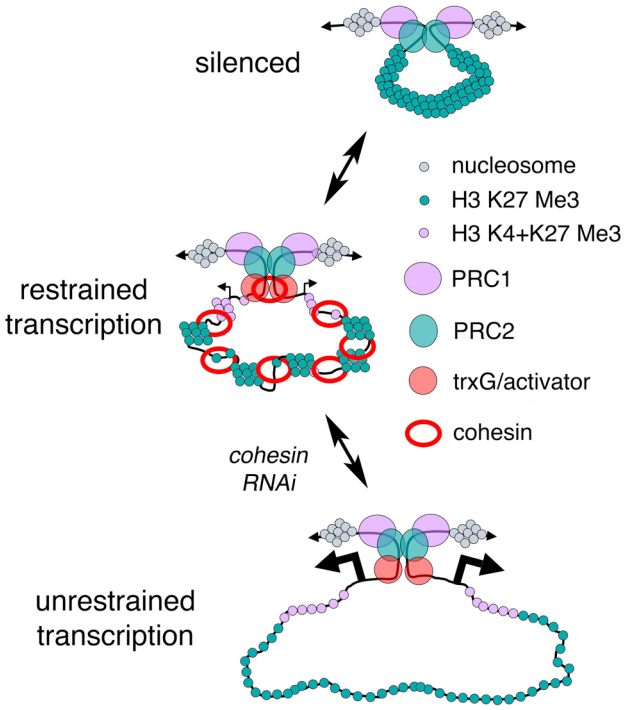

DAY 6
HLHm3
HLHmδ
engrailed
HLHm7
invected
HLHmγ
invected
HLHmβ
m2
10,885
log2 Nipped-B / Mock
Frequency
log2 Rad21 / Mock
6
4
2
0
-2
-4
-6
2000
1500
1000
500
0
-6
-4
-2
0
2
4
6
8

Figure A1.8


silenced
nucleosome
H3 K27 Me3
H3 K4+K27 Me3
PRC1
PRC2
trxG/activator
cohesin
restrained
transcription
cohesin
RNAi
unrestrained
transcription

Supplementary Tables

Table A1.S1. Regions of cohesin – H3K27Me3 overlap.

Region[a]	Size (kb)	Genes	Cell Type	Effect of cohesin RNAi[b]
2L 14,531,380 – 14,586,447	55.1	-	BG3	-
2R 2,390,180 – 2,434,731	44.6	*jing*	BG3	1.4-fold increase
2R 7,361,191 – 7,416,761	55.6	*invected*, *engrailed*	BG3	24-fold increase
2R 8,849,716 – 8,930,632	80.9	*Psc, Su(z)2*	BG3, Sg4	2-fold increase
3R 6,445,227 – 6,450,047	4.8	*hth*	BG3	4-fold increase
3R 11,472,282 – 11,481,587	9.3	-	BG3	-
3R 21,822,901 – 21,866,784	43.9	E(spl)-C	BG3	235-fold increase
3R 26,597,168 – 26,602,515	5.3	*zfh1*	Sg4	-
X 8,662,774 – 8,696,222	33.4	*Lim1*	BG3	2-fold increase

[a]Regions of overlap > 2 kb at $p \leq 10^{-3}$ with product of MAT scores > 50; positions are indicated by chromosome arm and nucleotide numbers from the April 2006 release of the *Drosophila* genome sequence.

[b]Maximal effect on transcript levels for any gene in region seen in BG3 expression microarray analysis with Rad21 or Nipped-B RNAi knockdown.

Table A2.S2. Half-lives of E(spl)-C transcripts.

Gene	Rad21/Mock[a]	Mock $t_{1/2}$ (min)[b]	Rad21 $t_{1/2}$ (min)[c]
HLHmδ	10.2	24.5	21.4
HLHmγ	1.9	16.7	18.2
mα	1.3	15.6	15.5
HLHm3	1.7	23.4	24.8
HLHm7	6.3	60	65

[a]Fold-increase in transcript level in Rad21 RNAi-treated cells 3 days after treatment.

[b]Half-life of transcript in mock-treated cells after Actinomycin D.

[c]Half-life of transcript in Rad21 RNAi-treated cells after Actinomycin D.

Table A1.S3. Effects of Rad21 and Nipped-B RNAi on precocious sister chromatid separation (PSCS) and hyperploidy.

		Hyperploid				Chromosomes			
Cell	RNAi	# Cells	#[a]	%	p value[b]	#[c]	# PSCS	% PSCS	p value PSCS[d]
BG3	Mock	71	3	4.2	--	322	28	9	--
BG3	Rad21	30	1	3.3	0.76	146	35	24	$1.4x10^{-5}$
BG3	Nipped-B	31	1	3.2	0.77	133	43	32	$1.7x10^{-9}$
Sg4	Mock	74	3	4.1	--	610	142	23	--
Sg4	Rad21	37	2	5.4	0.54	249	118	47	$6.6x10^{-12}$
Sg4	Nipped-B	37	4	11	0.17	310	94	30	$1.3x10^{-2}$

[a]BG3 cells are diploid male with four large autosomes, two 4th dot chromosomes, one X and one Y chromosome. Sg4 cells are partially tetraploid, with eight large autosomes, two 4th dot chromosomes, and two X chromosomes. Cells with one or more extra large chromosomes were scored as hyperploid.

[b]Comparison of RNAi treated to Mock by Fisher's exact test.

[c]Only X chromosomes and large autosomes with clear morphology were scored for PSCS.

[d]Comparison of RNAi treated to Mock by Fisher's exact test.

Table A1.S4. Effects of Rad21 and Nipped-B knockdown on gene expression in BG3 cells.

Too large to reproduce. Please see supplementary materials online at doi:10.1371/journal.pone.0006202.

Table A1.S5. RNA polymerase II and cohesin binding to genes that increase or decrease in expression with Rad21 or Nipped-B RNAi.

Expression Change[a]	Genes (G)[b]	PolIII (P)[c]	P ---- G	cohesin (C)[d]	C ----- G	P+C[e]	(P+C) ------- C	(P+C) ------- P
All		4282				816		0.19
Increase	333	225	0.68	189	0.57	157	0.83	0.70
Decrease	407	268	0.66	146	0.36	120	0.82	0.45
Increase vs Decrease[f]			0.34		9.7×10^{-9}			
Increase vs All[g]								2.4×10^{-57}
Decrease vs All[g]								2.3×10^{-20}

[a]All genes, or genes that increase or decrease in expression ≥ 2-fold in two or more RNAi treatments

[b]Number of genes (G) with indicated expression change

[c]Number of genes with indicated expression change that bind RNA polymerase II (PolII, P)

[d]Number of genes with indicated expression change that bind cohesin and Nipped-B (C)

[e]Number of genes with indicated expression change that bind both PolII (P) and cohesin (C)

[f]Comparison for PolII or cohesin binding with Fisher's exact test

[g]Comparison of PolII-binding genes for cohesin binding with Fisher's exact test

Table A1.S6. Gene Ontology Categories Affected by Cohesin and Nipped-B.

Too large to reproduce. Please see supplementary materials online at doi:10.1371/journal.pone.0006202.

Table A1.S7. PCR primers for making RNAi templates.

Target	Direction	Sequence
Nipped-B	Forward	TAATACGACTCACTATAGGGAGATTCGCTGTTGGGAACTATGCTGG
	Reverse	TAATACGACTCACTATAGGGAGATGTCGGTATCACTTTCATCGCACG
Nipped-B	Forward	TAATACGACTCACTATAGGGAGAGTTCAATAGCCAACGACGCCG
	Reverse	TAATACGACTCACTATAGGGAGATGGTCCACGACTCGCATAACCTC
Rad21	Forward	TAATACGACTCACTATAGGGAGACTGGTTGGCAGCACATTGGG
	Reverse	TAATACGACTCACTATAGGGAGAGCATATCAGCATGGGCGTCC
Rad21	Forward	TAATACGACTCACTATAGGGAGATGGGTGACGATTTTAATCAAGGAG
	Reverse	TAATACGACTCACTATAGGGAGACGCCTGTTTTCTGGAATTTCCTG
SA	Forward	TAATACGACTCACTATAGGGAGAGGGACACCACGAGCGGATA
	Reverse	TAATACGACTCACTATAGGGAGAGCCGTCATCAACTTCATGGC
SA	Forward	TAATACGACTCACTATAGGGAGATGACGCTCCTTTTGAGCCTG
	Reverse	TAATACGACTCACTATAGGGAGATCTCGACGTTTACGTGTGTAGGC
Pc	Forward	TAATACGACTCACTATAGGGCAAAGCCGAGGTGCTCAAG
	Reverse	TAATACGACTCACTATAGGGACGAATCGCCTTTCATGTCG
CP190	Forward	TAATACGACTCACTATAGGGAGATAAACGGACGACCCATTAGC
	Reverse	TAATACGACTCACTATAGGGAGATTATGTCCGAAAGGATTCGC

Table A1.S8. Primers for RT-PCR.

Gene	Direction	Sequence
invected	Forward	TTGGTCGGCGGTTCGTAACAGC
	Reverse	TGGGTTGGGTGATAAACTTGTCG
engrailed	Forward	TTCCACAATCAGACGCACACC
	Reverse	CGTATCATCCACATCCACATCAATG
Abd-B	Forward	TCCGCAAACAAGAAGACACACTCC
	Reverse	GGTATCAAAGGACACGACACGACG
HLHmg	Forward	AATCAACAAGTGCCTGGACGAG
	Reverse	GCAAATGGGTGACGGTAAGTTC
HLHmd	Forward	GCAAATGGGTGACGGTAAGTTC
	Reverse	TCCTTGAGTTCGTCCAGATACAGG
HLHmb	Forward	CACAGAGTCTCCGAGTCCGAATC
	Reverse	CCAGAACCATTTTGTTGTAGTTTGG
ma	Forward	GGAGGACGAGGAGGATGTCTATG
	Reverse	GACTGGCTGAAGGTTGGTGGTC
m1	Forward	AGAACGCATTCGTCTGTAAAAACC
	Reverse	TGGGGCAAAAAGTTGGACAAGC
m2	Forward	CAAGTCAACGCCAGAGGAGTCTATC
	Reverse	CGCTGCTAATCAATGTGGGTGTG
HLHm3	Forward	AGGGAGTAGTGGCTGGTGTTGG
	Reverse	CTCATCGGTTTGCTGTGTCTGC
m4	Forward	ACCGTTCCCGTTCACTTCGTCC
	Reverse	ATAGCGATGGCGTTGGAGGTGCTG
HLHm5	Forward	TTGGACACCTTGAAGACCTTGG
	Reverse	CTGCTGCTTGACGACCTGTTTG
m6	Forward	CCGACAGTCAGCGATACGATAGC
	Reverse	CCTCCAATCCCACTTGAGTTGC
HLHm7	Forward	AGTGGATGTGGCTTTTGGAACC
	Reverse	GACGATACTGAGTGGAGTGTTGACG
E(spl)m8	Forward	ATGAACAAGTGCCTGGACAACC
	Reverse	CTTCCTGAGCCACCTTCTTTGG
RpL32	Forward	ATCGGTTACGGATCGAACAAGC
	Reverse	GTTCTGCATGAGCAGGACCTCC

Appendix 2:The AP-1 transcription factor *Batf* controls TH17 differentiation[3]

3 This chapter was adapted from: Schraml, B. U., Hildner, K., Ise, W., Lee, W. L., Smith, W. A., Solomon, B., **Sahota, G.**, Sim, J., Mukasa, R., Cemerski, S., Hatton, R. D., Stormo, G. D., Weaver, C. T., Russell, J. H., Murphy, T. L. & Murphy, K. M. The AP-1 transcription factor Batf controls T(H)17 differentiation. *Nature* **460**, 405-9 (2009). I had initially tried to analyze some of the previous microarray datasets using PhyloCon. In this paper, we had EMSA data, so we analyzed using CONSENSUS. A prototype gibbs-based gapped motif finder was also built to see if the half-sites were separated by a variable gap (no significant variably gapped motifs were found).

Abstract

Activator protein 1 (AP-1) transcription factors are dimers of Jun, Fos, MAF and activating transcription factor (ATF) family proteins characterized by basic region and leucine zipper domains[1]. Many AP-1 proteins contain defined transcriptional activation domains (TADs), but *Batf* and the closely related *Batf3* (refs 2, 3) contain only a basic region and leucine zipper and have been considered inhibitors of AP-1 activity[3-8]. Here we show that *Batf* is required for the differentiation of IL-17-producing T helper (TH17) cells[9]. TH17 cells comprise a $CD4^+$ T cell subset that coordinates inflammatory responses in host defense but is pathogenic in autoimmunity[10-13]. *Batf*$^{-/-}$ mice have normal TH1 and TH2 differentiation, but show a defect in TH17 differentiation, and are resistant to experimental autoimmune encephalomyelitis (EAE). *Batf*$^{-/-}$ T cells fail to induce known factors required for TH17 differentiation, such as RORγt[11] and the cytokine IL-21 (refs 14-17). Neither addition of IL-21 nor overexpression of RORγt fully restores IL-17 production in *Batf*$^{-/-}$ T cells. The IL-17 promoter is Batf-responsive, and upon TH17 differentiation, Batf binds conserved intergenic elements in the *IL-17A/F* locus and to the IL-17, IL-21 and IL-22 (ref 18) promoters. These results demonstrate that the AP-1 protein Batf plays a critical role in TH17 differentiation.

Results and Discussion

In a gene expression survey (Supplementary Fig. A2.S1a), we identified the basic leucine zipper (bZIP) transcription factor ATF-like[7] (*Batf*) to be highly expressed in TH1,

TH2 and TH17 cells compared to naïve T cells and B cells. *Batf* and *Batf3* (refs 2, 3) form heterodimers with Jun[6,7] and are considered repressors of AP-1 activity[3,5,6,8,19]. To assess its role in T cell differentiation[20], we generated *Batf*$^{-/-}$ mice (Supplementary Fig. A2.S2a, b). *Batf*$^{-/-}$ mice lacked detectable *Batf* protein, were fertile and appeared healthy. Batf protein was low in naïve T cells, increased in TH2 cells, induced by activation (Supplementary Fig. A2.S2), present in the nucleus and cytoplasm, but upon activation showed increased nuclear translocation (Fig. A2.1a and Supplementary Fig. A2.S1b, c). *Batf*$^{-/-}$mice had normal thymus, spleen and lymph node development and CD4^{+} and CD8^{+} T cell development (Supplementary Figs. A2.S3, A2.S4a, b). Although *Batf*-transgenic mice had altered NKT cell development[21], *Batf*$^{-/-}$mice had normal development of NKT cells (Supplementary Fig. A2.S4c), B cells (Supplementary Fig. A2.S4d, e), conventional and plasmacytoid dendritic cells (Supplementary Fig. A2.S5a, b).

Batf$^{-/-}$ T cells displayed normal TH1 and TH2 differentiation (Supplementary Fig. A2.S6a). Under TH17 conditions, *Batf*$^{-/-}$T cells, but not *Batf*$^{+/-}$ T cells, showed a dramatic reduction in IL-17 production, but had normal levels of IL-2, IFN-γ and IL-10 (Fig. A2.1b, c). *Batf*$^{-/-}$DO11.10^{+} T cells showed loss of IL-17 even after several passages under TH17 conditions (Supplementary Fig. A2.S6b). *Batf*$^{-/-}$CD8^{+} T cells also failed to produce IL-17 (Supplementary Fig. A2.S6c). We generated transgenic mice expressing FLAG-tagged *Batf* under the control of the CD2 promoter[22]. *Batf*-transgenic DO11.10^{+} CD4^{+} T cells and CD8^{+} T cells had increased IL-17 production under TH17 conditions compared to controls (Supplementary Fig. A2.S6d, e). Lamina propria CD4^{+} T cells, which constitutively express IL-17 in wild type mice[11], failed to produce IL-17 in *Batf*$^{-/-}$ mice (Supplementary Fig. A2.S6f).

TH17 cells are the major pathogenic population in experimental autoimmune encephalomyelitis[10] (EAE), although factors other than IL-17A and IL-17F can contribute to disease[23]. *Batf*$^{+/+}$ mice immunized with myelin oligodendrocyte glycoprotein peptide 35-55 (MOG35-55) (Fig. A2.2) developed EAE, but *Batf*$^{-/-}$ mice were completely resistant (Fig. A2.2a). At peak disease, CNS-infiltrating and splenic CD4$^+$ T cells from *Batf*$^{+/+}$ mice produced abundant IL-17 and IFN-γ, while T cells from *Batf*$^{-/-}$mice produced no IL-17 (Fig. A2.2b, Supplementary Fig. A2.S7a). Since IL-6-deficient mice are resistant to EAE due to a compensatory increase in Foxp3$^+$ T regulatory (Treg) cells[14], we analyzed splenic *Batf*$^{+/+}$ and *Batf*$^{-/-}$CD4$^+$ T cells for Foxp3 expression before and after MOG35-55 immunization (Supplementary Fig. A2.S7b, c). *Batf*$^{-/-}$ mice had lower basal numbers of splenic Foxp3$^+$ T cells compared to *Batf*$^{+/+}$ mice, but showed no change in Foxp3$^+$ expression after MOG35-55 immunization (Supplementary Fig. A2.S7b, c), suggesting that their resistance to EAE is not due to an increase in Treg cells. To determine whether the resistance to EAE in *Batf*$^{-/-}$ mice resulted from a defect within T cells or other immune cells, we injected naïve *Batf*$^{+/+}$ CD4$^+$ T cells or PBS control buffer into mice before MOG35-55 immunization (Fig. A2.2c). *Batf*$^{-/-}$ mice receiving PBS remained resistant to EAE, but *Batf*$^{-/-}$ mice receiving naïve *Batf*$^{+/+}$ CD4$^+$ T cells developed severe EAE (Fig. A2.2c, Supplementary Table A2.S1) with CNS-infiltrating IL-17-producing CD4$^+$ T cells (Supplementary Fig. A2.S7d). Thus, *Batf*$^{-/-}$ mice have a T cell-intrinsic defect preventing EAE.

Batf could control TH17 development by regulating IL-6 or TGF-β signaling. IL-6 receptor expression and IL-6-induced STAT3 phosphorylation were normal in *Batf*$^{-/-}$T cells (Supplementary Fig. A2.S8a and b). TGF-β induced normal levels of Foxp3 in

Batf$^{-/-}$CD4^{+} T cells (Supplementary Fig. A2.S8d). While *Batf*$^{-/-}$T cells failed to fully downregulate Foxp3 in response to IL-6 (ref 12), neutralization of IL-2 abrogated increased Foxp3 in *Batf*$^{-/-}$ T cells, without restoring IL-17 production (Supplementary Fig. A2.S8d, e). Thus, *Batf*$^{-/-}$ T cells exhibit normal TGF-β signaling and proximal IL-6 signaling, implying *Batf* may regulate downstream target genes.

IL-21, an early target of IL-6 signaling in CD4^{+} T cells[17], is required for TH17 development[14-16]. IL-21 was reduced in *Batf*$^{-/-}$CD4^{+} T cells activated under TH17 conditions (Fig. A2.3a). Addition of IL-21 failed to rescue TH17 development in *Batf*$^{-/-}$T cells (Fig. A2.3b) but IL-21-induced STAT3 phosphorylation was intact (Supplementary Fig. A2.8c), suggesting that *Batf* regulates other factors besides IL-21 during TH17 differentiation.

We performed DNA microarrays and quantitative RT-PCR (qRT-PCR) of *Batf*$^{+/+}$ and *Batf*$^{-/-}$ T cells activated with combinations of IL-6 and/or TGF-β (Fig. A2.3c, d and Supplementary Fig. A2.S9). This analysis identified several genes known to regulate TH17 development as *Batf*-dependent (Fig. A2.3c, d, Supplementary Fig. A2.S9c and Supplementary Table A2.S2), including RORγt[17], RORα[24], the aryl hydrocarbon receptor[25,26], IL-22 (ref 18) and IL-17. However, IRF-4 (ref 13) and SOCS gene expression were unchanged in *Batf*$^{-/-}$ T cells (Supplementary Fig. A2.S9b and Supplementary Table A2.S4). Early induction of RORγt was normal in *Batf*$^{-/-}$ T cells but was not maintained at 62h after stimulation (Supplementary Fig. A2.S11a). *Batf* appeared necessary for expression of a subset of IL-6-induced genes, but was not required for expression of TGF-β-induced genes (Fig. A2.3c, Supplementary Fig. A2.S9a and Supplementary Table A2.S2, A2.S3). However, *Batf* did not globally affect IL-6-

induced responses, since IL-6-induced liver acute phase responses appeared normal in *Batf*$^{-/-}$ mice (Supplementary Fig. A2.S10).

Since RORγt acts directly on the IL-17 promoter[27,28], we asked whether RORγt could rescue TH17 development in *Batf*$^{-/-}$T cells. In *Batf*$^{+/+}$ T cells, retroviral RORγt expression induced 38% IL-17 production, compared to only 1.6% IL-17 production induced by control retrovirus (Fig. A2.3e and Supplementary Fig. A2.S11c)[11,13]. But in *Batf*$^{-/-}$T cells, retroviral RORγt expression induced only 5.7% IL-17 production (Fig. A2.3e and Supplementary Fig. A2.S11c). Even under TH17-inducing conditions, retroviral RORγt expression did not fully restore IL-17 production in *Batf*$^{-/-}$ T cells (Supplementary Fig. A2.S11b, c). Retroviral expression of both Batf and RORγt in *Batf*$^{-/-}$ T cells induced 26% IL-17 production, compared to only 5% with RORγt alone, and 14% with Batf alone (Supplementary Fig. A2.S11d), suggesting potential synergy between RORγt and Batf, and a possible direct action of Batf in transcription of IL-17 and other TH17-specific genes.

We used a reverse-strand retroviral reporter[29] to examine IL-17 promoter activity in primary *Batf*$^{+/+}$ and *Batf*$^{-/-}$ T cells (Fig. A2.4a). Three days after activation, *Batf*$^{-/-}$CD4^{+} T cells showed considerably less reporter activity than *Batf*$^{+/+}$ T cells, suggesting the proximal IL-17 promoter is *Batf*-responsive (Fig. A2.4a). Using chromatin immunoprecipitation (ChIP) analysis of several conserved regions within the *IL-17a/IL-17f* locus (Supplementary Fig. A2.S12a), we found that Batf specifically bound to the +9.6kb and +28kb intergenic regions within 24h after activation (Fig. A2.4b, Supplementary Fig. A2.S12b, c). By day 5 after stimulation, Batf bound specifically to several intergenic regions and to the proximal *IL-17a* and *IL-17f* promoters (Fig. A2.4b,

Supplementary Fig. A2.S12b, c), with distal elements showing more rapid and stronger binding than proximal elements.

We next examined Batf binding to a consensus AP-1 probe[6] by EMSA. This probe formed two complexes in *Batf*$^{+/+}$ TH17 cell extracts (Fig. A2.4c) that were dependent on stimulation (Supplementary Fig. A2.S13a). Only the upper complex formed in *Batf*$^{-/-}$ TH17 cells (Fig. A2.4c). An anti-Batf antibody inhibited the lower complex. In CD2-N-FLAG-*Batf*-transgenic TH17 cell extracts, the lower complex was specifically supershifted by an anti-FLAG antibody (Fig. A2.4c). Thus, only the lower complex binding the consensus AP-1 probe in TH17 cells contains Batf.

Several potential Batf binding sites were identified by EMSA in the IL-17, IL-21 and IL-22 proximal promoters, including the IL-17 promoter region (-188 to -210) that bound Batf in ChIP (Fig. A2.4b, Supplementary Fig. A2.S13b-d). Another Batf-binding IL-17 promoter region (-155 to -187) overlapped with a reported RORγt-binding element[27]. As an EMSA probe, this region forms two complexes in TH17 cells (Fig. A2.4d), with the lower complex being selectively inhibited by anti-Batf antibody, absent in *Batf*$^{-/-}$TH17 cells, and supershifted by an anti-FLAG antibody in *Batf*-transgenic TH17 extracts (Fig. A2.4d). We confirmed Batf binding to the IL-21 and IL-22 promoters by ChIP analysis (Supplementary Fig. A2.S13e). The program CONSENSUS[30] determined that the Batf-binding element in the IL-17, IL-21 and IL-22 promoters resembles canonical AP-1 elements at positions 1 through 3, with variation at remaining nucleotides (Supplementary Fig. A2.S13f). CONSENSUS did not identify other transcription factor binding sites enriched near Batf binding elements. We determined the composition of the Batf-containing complex using supershift analysis (Fig. A2.4e). The upper complex

supershifted with pan-anti-Fos antibody, whereas the lower complex supershifted with a pan-anti-Jun and anti-Batf antibodies. Anti-JunB supershifted the majority of the lower complex, but antibodies to c-Jun, JunD, ATF1 or ATF3 did not. Thus, Batf forms heterodimers preferentially with JunB during TH17 differentiation.

Although *Batf* and *Batf3* were considered AP-1 inhibitors[3-8], we have shown that they are required for the development of specific immune lineages[2]. *Batf* is selectively required for TH17 development, but unlike *Irf4* (Ref 13), is not required for TH2 development. Since *Batf* is also expressed in TH1 and TH2 cells, it likely cooperates with other TH17-specific factors to regulate target genes. Future work will determine whether the actions of *Batf* involve distinct DNA binding specificity or unique protein-protein interactions with TH17 specific factors.

Acknowledgements

We thank Dr. Roger Lallone (Brookwood Biomedical) for anti-Batf antibody preparation, and Dr. Barry Sleckman for Cre-expressing adenovirus. This work was supported by the Howard Hughes Medical Institute (KMM), and grants from the NIH HG00249 and training grant GM07200 (GDS), AI035783 (CTW), AR049293 (RDH) and from Daiichi-Sankyo Co. Ltd. (CTW).

Author Contributions

BUS generated *Batf*$^{-/-}$ mice, designed and analyzed the experiments, interpreted results and wrote the manuscript. KH constructed the targeting vector and probes, transgenic vector, and recombinant Batf. WI helped with retroviral expression experiments. WLL helped with reverse-strand reporter analysis. WAES helped with

mouse generation. BS helped with EMSA analysis. GS and GDS performed bioinformatics analysis for the Batf binding elements. JS and JHR helped with EAE experiments. RM, RDH and CTW performed ChIP experiments. TLM and SC performed confocal microscopy for Batf. KMM directed the study and wrote the manuscript.

Author information

Microarray data are available at Array Express, E-MEXP-1518, E-MEXP-2152 and E-MEXP-2153. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to KMM (kmurphy@wustl.edu).

Methods Summary

Mice

Batf$^{-/-}$ mice were generated by homologous recombination, deleting exons 1 and 2 of the *Batf* gene on the pure 129SvEv genetic background. The neomycin resistance cassette was removed from the targeted *Batf* allele in ES cells before generation of mice.

T cell differentiation assays

Naïve $CD4^+CD62L^+CD25^-$ T cells were isolated by cell sorting and activated with plate-bound anti-CD3 and soluble anti-CD28 antibodies. Cultures were supplemented with anti-IL-4 (11B11; hybridoma supernatant), IFN-γ (Peprotech; 0.1ng/ml) and IL-12 (Genetics Institute; 10U/ml) for TH1; anti-IFN-γ (H22; BioXcell; 10μg/ml), anti-IL-12 (Tosh; BioXcell; 10μg/ml) and IL-4 (Peprotech; 10ng/ml) for TH2; anti-IL-4, anti-IL-12, anti-IFN-γ, IL-6 (Peprotech 20ng/ml) and TGF-β (Peprotech; 0.5ng/ml) for TH17

differentiation. Unless otherwise indicated, three days after activation cells were restimulated with PMA/ionomycin for 4h for intracellular cytokine analysis by flow cytometry.

Intracellular Staining

For intracellular cytokine staining, cells were stained for surface markers followed by fixation with 2% formaldehyde for 15 minutes at room temperature.
Cells were then washed once in 0.05% saponin and stained with anti-cytokine antibodies in 0.5% saponin. Anti-phospho-STAT3 antibody (BD Pharmingen) was used according to the manufacturer's recommendations. Briefly, cells were stained for surface markers followed by fixation with 90% methanol at -20^{o}C overnight. Cells were then washed and stained for phospho-Stat3 in PBS containing 3% FCS. Foxp3 staining was performed according to the manufacturer's recommendations using Foxp3 staining buffers (eBioscience).

Induction of EAE

Mice (7-10 weeks old) were immunized subcutaneously with 100µg MOG35-55 peptide (Sigma) emulsified in CFA (IFA supplemented with 500µg *Mycobacterium tuberculosis*). One and three days later mice were given 300ng Pertussis Toxin (List Biological Laboratories) intraperitonally (i.p.). Clinical scores were assessed as described in methods. For T cell transfer experiments mice were injected with either PBS or 10^7 *Batf*$^{+/+}$ CD4$^+$ T cells 4 days prior to MOG35-55 immunization[13].

References

1. Wagner,E.F. & Eferl,R. Fos/AP-1 proteins in bone and the immune system. *Immunol Rev* **208**, 126-140 (2005).

2. Hildner,K. *et al.* Batf3 deficiency reveals a critical role for CD8alpha+ dendritic cells in cytotoxic T cell immunity. *Science* **322**, 1097-1100 (2008).

3. Iacobelli,M., Wachsman,W. & McGuire,K.L. Repression of IL-2 promoter activity by the novel basic leucine zipper p21SNFT protein. *J Immunol* **165**, 860-868 (2000).

4. Blank,V. Small Maf proteins in mammalian gene control: mere dimerization partners or dynamic transcriptional regulators? *J Mol Biol* **376**, 913-925 (2008).

5. Williams,K.L. *et al.* Characterization of murine BATF: a negative regulator of activator protein-1 activity in the thymus. *Eur. J Immunol* **31**, 1620-1627 (2001).

6. Echlin,D.R., Tae,H.J., Mitin,N. & Taparowsky,E.J. B-ATF functions as a negative regulator of AP-1 mediated transcription and blocks cellular transformation by Ras and Fos. *Oncogene* **19**, 1752-1763 (2000).

7. Dorsey,M.J. *et al.* B-ATF: a novel human bZIP protein that associates with members of the AP-1 transcription factor family. *Oncogene* **11**, 2255-2265 (1995).

8. Thornton,T.M., Zullo,A.J., Williams,K.L. & Taparowsky,E.J. Direct manipulation of activator protein-1 controls thymocyte proliferation in vitro. *Eur. J Immunol* **36**, 160-169 (2006).

9. Harrington,L.E. *et al.* Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nat Immunol* **6**, 1123-1132 (2005).

10. Langrish,C.L. *et al.* IL-23 drives a pathogenic T cell population that induces

autoimmune inflammation. *J Exp. Med.* **201**, 233-240 (2005).

11. Ivanov,I.I. *et al.* The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells. *Cell* **126**, 1121-1133 (2006).

12. Bettelli,E. *et al.* Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells. *Nature* **441**, 235-238 (2006).

13. Brustle,A. *et al.* The development of inflammatory T(H)-17 cells requires interferon-regulatory factor 4. *Nat Immunol* **8**, 958-966 (2007).

14. Korn,T. *et al.* IL-21 initiates an alternative pathway to induce proinflammatory T(H)17 cells. *Nature* **448**, 484-487 (2007).

15. Nurieva,R. *et al.* Essential autocrine regulation by IL-21 in the generation of inflammatory T cells. *Nature* **448**, 480-483 (2007).

16. Wei,L., Laurence,A., Elias,K.M. & O'Shea,J.J. IL-21 is produced by Th17 cells and drives IL-17 production in a STAT3-dependent manner. *J Biol Chem.* **282**, 34605-34610 (2007).

17. Zhou,L. *et al.* IL-6 programs T(H)-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nat Immunol* **8**, 967-974 (2007).

18. Liang,S.C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J Exp. Med.* **203**, 2271-2279 (2006).

19. Bower,K.E., Fritz,J.M. & McGuire,K.L. Transcriptional repression of MMP-1 by p21SNFT and reduced in vitro invasiveness of hepatocarcinoma cells. *Oncogene* **23**, 8805-8814 (2004).

20. Hess,J., Angel,P. & Schorpp-Kistner,M. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci* **117**, 5965-5973 (2004).

21. Williams,K.L. *et al.* BATF transgenic mice reveal a role for activator protein-1 in NKT cell development. *J Immunol* **170**, 2417-2426 (2003).

22. Zhumabekov,T., Corbella,P., Tolaini,M. & Kioussis,D. Improved version of a human CD2 minigene based vector for T cell-specific expression in transgenic mice. *J Immunol Methods* **185**, 133-140 (1995).

23. Haak,S. *et al.* IL-17A and IL-17F do not contribute vitally to autoimmune neuro-inflammation in mice. *J Clin. Invest* **119**, 61-69 (2009).

24. Yang,X.O. *et al.* T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. *Immunity* **28**, 29-39 (2008).

25. Veldhoen,M. *et al.* The aryl hydrocarbon receptor links T(H)17-cell-mediated autoimmunity to environmental toxins. *Nature* (2008).

26. Quintana,F.J. *et al.* Control of T(reg) and T(H)17 cell differentiation by the aryl hydrocarbon receptor. *Nature* **453**, 65-71 (2008).

27. Ichiyama,K. *et al.* Foxp3 inhibits RORgammat-mediated IL-17A mRNA transcription through direct interaction with RORgammat. *J Biol Chem.* **283**, 17003-17008 (2008).

28. Zhang,F., Meng,G. & Strober,W. Interactions among the transcription factors Runx1, RORgammat and Foxp3 regulate the differentiation of interleukin 17-producing T cells. *Nat Immunol* **9**, 1297-1306 (2008).

29. Zhu,H. *et al.* Unexpected characteristics of the IFN-gamma reporters in nontransformed T cells. *J. Immunol.* **167**, 855-865 (2001).

30. Hertz,G.Z. & Stormo,G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* **15**, 563-577 (1999).

Figure legends.

Figure A2.1. Loss of IL-17 production in *Batf*$^{-/-}$ T cells. a, DO11.10^{+}CD4^{+} T cells from CD2-N-FLAG-*Batf* transgenic mice or littermates were cultured with OVA/APCs under TH2 conditions for 7 days, and stained with antibodies to CD4 and FLAG. **b,** *Batf*$^{+/+}$ and *Batf*$^{-/-}$CD4^{+}CD62L^{+}CD25^{-}T cells cultured under TH17 conditions were restimulated with PMA/ionomycin on days 7 (left panel) or 3 (middle and right panels) and stained for IL-17, IFN-γ, IL-2 and IL-10. **c,** IL-17 and IFN-γ expression in DO11.10^{+}CD4^{+} T cells from *Batf*$^{+/+}$, *Batf*$^{+/-}$and *Batf*$^{-/-}$ mice activated with OVA/APCs under TH17 conditions. Data are representative of at least 2 independent experiments.

Figure A2.2. *Batf*$^{-/-}$ mice are resistant to EAE. a, *Batf*$^{+/+}$ (n=12) and *Batf*$^{-/-}$ (n=13) mice were immunized with MOG33-35 peptide. (Mean clinical EAE scores +/- s.e.m, representative of two independent experiments). **b,** 13 days after EAE induction, CNS-infiltrating lymphocytes were stimulated with PMA/ionomycin, gated on CD4^{+} cells and stained for intracellular IL-17 and IFNγ (Clinical scores are in parentheses, data are representative of 2-3 mice per group). **c,** *Batf*$^{+/+}$ and *Batf*$^{-/-}$mice were injected with control PBS buffer (n=5) or 1x10^{7} *Batf*$^{+/+}$ CD4^{+} T cells (n=6) four days prior to EAE induction. Mean clinical scores are shown.

Figure A2.3. Batf controls multiple TH17-associated genes. a, IL-21 expression in *Batf*$^{+/+}$ or *Batf*$^{-/-}$ T cells cultured under TH17 conditions determined by qRT-PCR and ELISA. (mean + s.d. 3 mice). **b,** IL-17 and IFN-γ expression of CD4^{+}CD62L^{+}CD25^{-}T cells cultured in **a** in the presence or absence of IL-21. **c,** Microarray analysis of anti-CD3/CD28-activated T cells at 72h, presented as heat maps of genes 5-fold-induced in *Batf*$^{+/+}$ T cells under TH17 conditions. **d,** IL-17 and IL-22 expression in *Batf*$^{+/+}$ or *Batf*$^{-/-}$

$CD4^+$ T cells activated under TH17 conditions for 3 days. **e,** Anti-CD3/CD28-activated *Batf*$^{+/+}$ or *Batf*$^{-/-}$$CD4^+$ T cells were left uninfected or infected with RORγt-GFP-RV or control-GFP-RV, and stained for IL-17.

Figure A2.4. Batf directly regulates IL-17 expression. a, *Batf*$^{+/+}$ and *Batf*$^{-/-}$ $CD4^+$ T cells cultured under TH17 conditions were infected with hCD4-pA-GFP-RV-IL-17p reporter virus. GFP expression after PMA/ionomycin restimulation is shown. **b,** *Batf*$^{+/+}$ and *Batf*$^{-/-}$$CD4^+$ T cells cultured under TH17 conditions for 5 days were subjected to ChIP analysis of the indicated regions using anti-Batf antibody (mean + s.d.). **c, d, f,** EMSA supershift analysis of TH17 whole cell extracts using a consensus AP-1 **(c, f)** or the IL-17(-155 to -187) probe **(d).** (*Batf*$^{+/+}$ (WT), *Batf*$^{-/-}$(KO), CD2-N-FLAG-*Batf* transgenic (TG), IL-17(-155 to -187) and RORE probes were used as competitors).

Figure A2.1

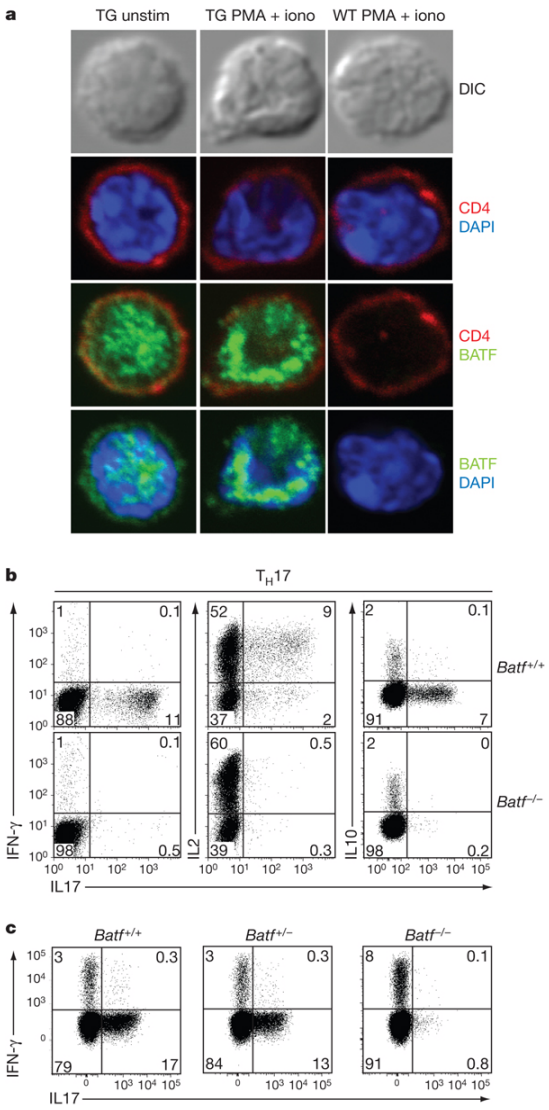

a
TG unstim
TG PMA + iono
WT PMA + iono
DIC
CD4
DAPI
CD4
BATF
BATF
DAPI
b
T_H17
IFN-γ
IL17
IL2
IL10
Batf+/+
Batf−/−
c
Batf+/+
Batf+/−
Batf−/−
IFN-γ
IL17

Figure A2.2

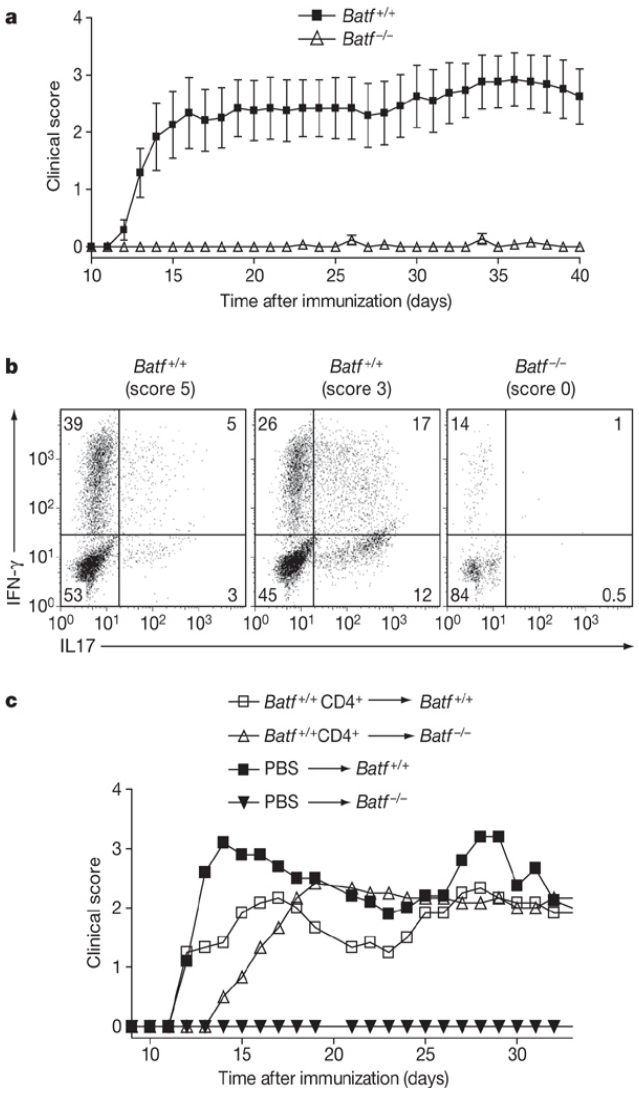

a
Batf+/+
Batf−/−
Clinical score
Time after immunization (days)
b
Batf+/+ (score 5)
Batf+/+ (score 3)
Batf−/− (score 0)
39
5
53
3
26
17
45
12
14
1
84
0.5
IFN-γ
IL17
c
Batf+/+ CD4+ → Batf+/+
Batf+/+CD4+ → Batf−/−
PBS → Batf+/+
PBS → Batf−/−
Clinical score
Time after immunization (days)

Figure A2.3

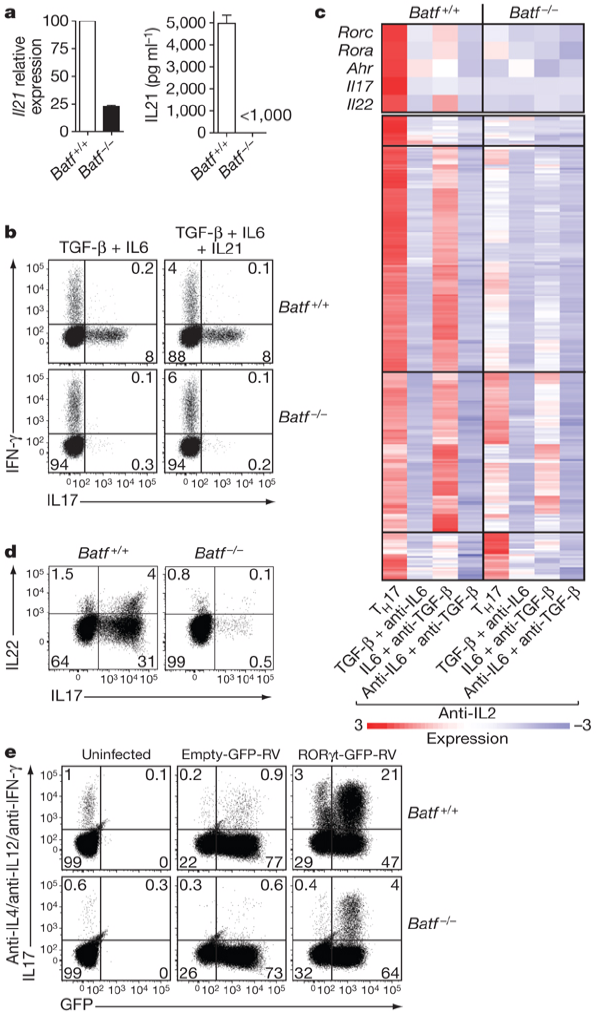
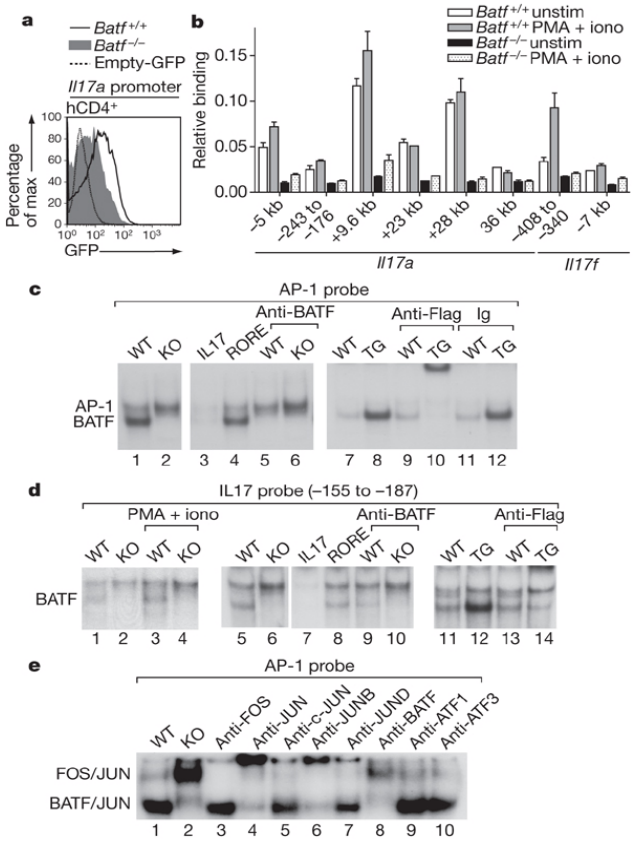

a
Il21 relative expression
100
75
50
25
0
Batf+/+
Batf−/−
IL21 (pg ml−1)
5,000
4,000
3,000
2,000
1,000
0
<1,000
b
TGF-β + IL6
TGF-β + IL6 + IL21
Batf+/+
Batf−/−
IFN-γ
IL17
c
Batf+/+
Batf−/−
Rorc
Rora
Ahr
Il17
Il22
TH17
TGF-β + anti-IL6
IL6 + anti-TGF-β
Anti-IL6 + anti-TGF-β
Anti-IL2
3
−3
Expression
d
Batf+/+
Batf−/−
IL22
IL17
e
Uninfected
Empty-GFP-RV
RORγt-GFP-RV
Batf+/+
Batf−/−
Anti-IL4/anti-IL12/anti-IFN-γ
IL17
GFP

Figure A2.4


a
Batf+/+
Batf−/−
Empty-GFP
Il17a promoter
hCD4+
Percentage of max
GFP
b
Batf+/+ unstim
Batf+/+ PMA + iono
Batf−/− unstim
Batf−/− PMA + iono
Relative binding
0.15
0.10
0.05
0.00
−5 kb
−243 to −176
+9.6 kb
+23 kb
+28 kb
36 kb
−408 to −340
−7 kb
Il17a
Il17f
c
AP-1 probe
Anti-BATF
Anti-Flag
Ig
WT KO IL17 RORE WT KO WT TG WT TG WT TG
AP-1
BATF
1 2 3 4 5 6 7 8 9 10 11 12
d
IL17 probe (−155 to −187)
PMA + iono
Anti-BATF
Anti-Flag
WT KO WT KO WT KO IL17 RORE WT KO WT TG WT TG
BATF
1 2 3 4 5 6 7 8 9 10 11 12 13 14
e
AP-1 probe
WT KO Anti-FOS Anti-JUN Anti-c-JUN Anti-JUNB Anti-JUND Anti-BATF Anti-ATF1 Anti-ATF3
FOS/JUN
BATF/JUN
1 2 3 4 5 6 7 8 9 10

Online Methods

Generation of *Batf* $^{-/-}$ mice.

Murine *Batf* exons 1–2 were deleted by homologous recombination via a targeting vector constructed in pLNTK[31] using a 1 kb genomic fragment (left arm) upstream of the *Batf* exon 1 and a 3.6 kb genomic fragment (right arm) downstream of exon 2. The left arm was generated by PCR from genomic DNA with the use of the following oligonucleotides: left arm forward (5'-ATTA**CTCGAG**TGAAACAAACAGGCAGTCGCAGTG) and left arm reverse (5'-ATTA**CTCGAG**CCTACTACCTTTCAGGGCTACTGC). The right arm was generated by PCR with the use of the following oligonucleotides: right arm forward (5'-ATTA**GTCGAC**GCATTCTTCATGGTCCTTAGCCTTGG) and right arm reverse (5'-ATTA**GTCGAC**CAGAGAATGAGAAATGTTGGAGG). EDJ22 embryonic stem cells were transfected with linearized targeting vector and targeted clones were identified by Southern blot analysis using probes A and B located 5' to the left arm and 3' to the right arm respectively. Probe A was generated using the oligonucleotides 5'-CAACTGGGTCTGAGTCAAGAGGT and 5'-CGTAGCCGCTGATTGTTTTAGAAC to generate a 531bp product. Probe B was generated using the oligonucleotides 5'-ACAGCTTGAACTTCAGAGCCCTCC and 5'-CACATTTAAGTCACAATAACACTGC to generate a 772bp product. The neomycin resistance cassette was deleted from successfully targeted clones by *in vitro* treatment with Adeno-Cre virus (gift from Dr. Barry Sleckman, Washington University, St. Louis, MO) and targeted clones with successful neo deletion were identified by Southern blot

using probes A and B (Supplementary Fig. A2.S2a and b). Blastocyst injections were performed with two distinct recombinant clones each of which generated germline transmission of the targeted *Batf* allele. Male chimeras were crossed with 129SvEv females to establish *Batf* mutants on the pure 129SvEv genetic background. All experiments were performed with mice harboring the neo-deleted mutant allele. Homozygous mice were obtained by intercrossing heterozygous siblings and littermates were used as controls in most experiments. For some experiments 129SvEv wild type mice purchased from Taconic served as controls. For experiments with DO11.10 transgenic Batf $^{-/-}$mice, mice were crossed to BALB/c mice for at least 5 generations and littermates were used as control. For the generation of transgenic mice, *Batf* cDNA was cloned from CD4^{+} T cell mRNA using primers 5'-GGAAGATTAGAACCATGCCTC and 5'-AGAAGGTCAGGGCTGGAAG and subcloned into the GFP-RV retrovirus[32]. An N-terminal FLAG tag was introduced by Quick Change Mutagenesis kit (Stratagene) using the primers 5'-

GGACTACAAAGACGATGACGACAAGCCTCACAGCTCCGACAGCA and 5'-CTTGTCGTCATCGTCTTTGTAGTCCATGGTTCTAATCTTCCAGATC. The underlined sequence indicates nucleotides used to introduce the FLAG-tag. The FLAG-tagged *Batf* was cloned into the CD2 microinjection cassette[33] via blunt end strategy into Sma1 digested CD2 microinjection cassette. Transgene expression in CD4^{+} T cells was tested by anti-FLAG western blot. CD2-N-FLAG-*Batf* transgenic mice were crossed to C57BL/6 and BALB/c mice for at least 5 generations. Transgene-negative littermates were used as control mice. Mice were bred and maintained at the animal facilities at Washington University in St. Louis. All animal experiments were approved by the

Animal Studies Committee at Washington University.

Visualization of lymph nodes.

To visualize superficial inguinal lymph nodes mice were injected with 50 l of 1% Evans Blue dye solution into each hind foot pad. After 1.5 hours mice were sacrificed and lymph nodes were visualized using a dissecting microscope[34].

Western Blot analysis.

To test for residual Batf protein expression, total splenocytes from *Batf*$^{++}$ and *Batf*$^{-/-}$ 129SvEv mice were stimulated with anti-CD3 for 3 days under TH17 conditions. Cells were then lysed in RIPA buffer, electrophoresed on 15% polyacrylamide gels, transferred to nitrocellulose and analyzed by Western Blot with rabbit anti-murine Batf polyclonal serum and HRP-conjugated anti-rabbit Ig antibody (Jackson ImmunoResearch). Affinity purified rabbit anti-murine Batf polyclonal serum (Brookwood Biomedical; Birmingham, AL) was generated by immunization with full length recombinant Batf protein. Equal protein loading was assessed by subsequent immunoblotting with antibody to -actin (Santa Cruz Biotechnology) and HRP conjugated anti-mouse antibody (Jackson ImmunoResearch). For analysis of Batf protein expression in naïve CD4$^+$ T cells, magnetically purified CD4$^+$ T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$129SvEv mice were isolated. Equal cell numbers were lysed in RIPA buffer and subjected to Western Blot analysis as described above. For analysis of Batf expression in TH2 cells, magnetically purified CD4$^+$ T cells from Batf^{++} and Batf$^{-/-}$ mice were activated with anti-CD3/CD28 in the presence of IL-4, anti-IL-12 (Tosh), and anti-IFN- (H22). On day 4 cells were left unstimulated or stimulated with PMA/ionomycin for 4 hrs. Cells were collected by

centrifugation, washed with PBS, and resuspended ($100e^6$ cells/ml) in Affymetrix Chip lysis buffer (10mM Tris pH 7.5, 10mM NaCl, 3mM MgCl2, 0.5% IGEPAL, with protease inhibitors (PMSF, aprotinin, leupeptin)). After 5 min at 4°C, nuclei were collected by centrifugation (800 rcf for 3 min 4°C) and lysed in RIPA ($100e^6$ cell equivalents/ml) with protease inhibitors. Nuclear lysates were centrifuged for 10 min 4°C 15000 rcf, diluted with an equal volume of 2x SDS-PAGE sample buffer containing 2-ME and extracts from equal cell numbers were subjected to Western Blot analysis using rabbit anti-murine Batf polyclonal serum. Equal protein loading was assessed by subsequent immunoblotting with antibody to Lamin B (Santa Cruz Biotechnology) and HRP conjugated anti-goat Ig (Jackson ImmunoResearch).

Immunohistochemistry.

$CD4^+$ T cells from CD2-N-FLAG-*Batf* transgenic mice were isolated by magnetic separation and either left untreated or stimulated with PMA/ionomycin for 4h. Cells were then allowed to settle on poly-L-lysine treated slides, fixed with 4% Formaldehyde, permeabilized with 0.25% TritonX100 and were stained with an anti-FLAG antibody (M2, Sigma Aldrich) according to the manufacturer's recommendations. A goat anti-mouse AF-488 (Invitrogen) antibody was to detect anti-FLAG staining. For analysis of cellular localization of Batf in TH2 cells, DO11.10 $CD4^+$ T cells from CD2-N-FLAG-*Batf* transgenic mice were isolated and differentiated with OVA and APC under TH2 conditions for 7 days. On day 7 cells were either left untreated or stimulated with PMA/ionomycin for 4h. Cells were stained with anti-FLAG antibody as described above. Cells were also stained with anti-CD4APC antibody (BDBiosciences). Confocal images were obtained with the Olympus FV1000 microscope and software using a 60x oil

objective. The pinhole was set to 110 m. The excitation/emission settings used for DAPI, Alexa 488 and Alexa 633 were 405/461nm, 488/520nm and 635/668nm respectively.

Additional methods can be found in the Supplementary Information.

Methods Reference List

31. Gorman,J.R. *et al.* The Ig(kappa) enhancer influences the ratio of Ig(kappa) versus Ig(lambda) B lymphocytes. *Immunity.* **5**, 241-252 (1996).

32. Ranganath,S. *et al.* GATA-3-dependent enhancer activity in IL-4 gene regulation. *J. Immunol.* **161**, 3822-3826 (1998).

33. Zhumabekov,T., Corbella,P., Tolaini,M. & Kioussis,D. Improved version of a human CD2 minigene based vector for T cell-specific expression in transgenic mice. *J Immunol Methods* **185**, 133-140 (1995).

34. Sun,Z. *et al.* Requirement for RORgamma in thymocyte survival and lymphoid organ development. *Science* **288**, 2369-2373 (2000).

Supplementary Tables

Supplementary Table A2.S1. Transfer of *Batf* $^{+/+}$ CD4^{+} T cells into *Batf* $^{-/-}$ mice restores EAE

Group	Incidence	Mean Max. Score	Mortality	Mean day of onset‡
PBS→*Batf* +/+	5 of 5 (100%)	3.4 ± 0.7	1of 5 (20%)	12±0.8¶
PBS→*Batf* -/	0 of 5 (0%)	0	0 of 13 (0%)	NA
Batf +/+CD4+→*Batf* +/+	5 of 6 (83%)	3.0 ± 0.6	0 of 6 (0%)	13.6±2.3§¶
Batf +/+CD4+→*Batf* -/-	4 of 6 (66%)	2.4 ± 1.0	2 of 6 (33%)	15.5±1.7§

Four days prior to induction of EAE mice were injected with 1x10^{7} CD4^{+}*Batf* $^{+/+}$ T cells or control buffer (PBS) as indicated. The mice were monitored for disease development as described in Methods. Mean maximum score of disease was calculated and is presented ± s.e.m. Only animals positive for disease are included in the analysis. § not significant (p=0.215). ¶ not significant (p=0.232). NA, not applicable.

Supplementary Table A2.S2. Microarray data accompanying Figure A2.3c.

Too large to reproduce. Please see supplementary materials online at

doi:10.1038/nature08114

Supplementary Table A2.S3. Microarray data accompanying Supplementary Figure A2.S9a.

Too large to reproduce. Please see supplementary materials online at doi:10.1038/nature08114

Supplementary Table A2.S4. Microarray data accompanying Supplementary Figure A2.S9b.

Too large to reproduce. Please see supplementary materials online at doi:10.1038/nature08114

Supplementary Table A2.S5. RT-PCR primers and probes.

5'→3' Primers (*5'FAM 3' BHQ1 Probes*)	Location
IL17a -97 **(-97kb)**	
AAATGTGAGCCCCAGATCGA	Chr1:20,623,606-20,623,625
CTGCTGCTGTCCCAGGCACAGTTG	*Chr1:20,623,627-20,623,650*
GGGACATTTTTTCCACCATGA	Chr1:20,623,652-20,623,672
IL17a -60 **(-60kb)**	
TTGTCCCCTGGCTGTTCCT	Chr1:20,661,177-20,661,247
CCTTATCCAGCTGTCTTTTTCTCT	*Chr1:20,661.249-20,661,272*
GGGCTCCCCAAAAATTCACA	Chr1:20,661,274-20,661,293
IL17a -37 **(-37kb)**	
GTCCCTCTGTTGTTTCCAAGGAT	Chr1:20,683,616-20,683,638
TCATTGAGTCCTTCCAGCAGAGATTTCAGG	*Chr1:20,683,640-20,683,669*
GCCATTTCAGCCACTGTGAA	Chr1:20,683,671-20,683,690
IL17a -15 **(-15kb)**	
TGGCAAATGTTTTGTCAACCA	Chr1:20,705,507-20,705,527
TTCCTCGATTGCTGTCTACTCATC	*Chr1: 20,705,529-20,705,552*
CATGCAGCCTCTGCTTGAGA	Chr1:20,705,554-20,705,573
IL17a -5 **(-5kb)**	
CGATACTTTTCAGTGACATCCGTTT	Chr1:20,715,852-20,715,876
ACTTGAAACCCAGTCAGTTGCTGACCTTGA	*Chr1:20,715,879-20,715,908*
TGCTGACTTCATCTGATACCCTTAGA	Chr1:20,715,910-20,715,935
IL17a promoter **(-243 to -176)**	
GAACTTCTGCCCTTCCCATCT	Chr1:20,720,800-20,720,820
CCTTCGAGACAGATGTTGCCCGTCA	*Chr1: 20,720,822-20,720,846*
CAGCACAGAACCACCCCTTT	Chr1:20,720,848-20,720,867
IL17a +9.6 **(+9.6kb)**	
ATTTAGGGCACAGGTGACATGA	Chr1:20,730,688-20,730,709
TGGTTCTCAAAGCATAAACCTCATTC	*Chr1:20,730,711-20,730,736*
CCACTTCCCCGACCTCACTA	Chr1:20,730,738-20,730,757
IL17a +23 **(+23kb)**	
CAAATCCGTGTGCCTTCTGTT	Chr1:20,744,816-20,744,836
CTGCAGTGAGGAAGATGTTTCCAATGAGG	*Chr1:20,744,838-20,744,866*
AGGTTGACTTCGTCCCTGTGA	Chr1:20,744,870-20,744,890
IL17a +28 **(+28kb)**	

GTGGCCTACTTCAGGCAGATG	Chr1:20,749,994-20,750,014
TGAGAAGCCAGCGTCGGGTCC	*Chr1:20,750,016-20,750,036*
GGAGCCGATGAGAAGCATTC	Chr1:20,750,039-20,750,058

IL17a +36 **(+36kb)**

AGATAATGTATCACACAGCCCTGAAG	Chr1:20,757,551-20,757,576
AGCCAGTGCCTTAATCCATTGGG	*Chr1:20,757,578-20,757,600*
CATGGTTGTGAAGTTGGTGAGATG	Chr1:20,757,602-20,757,625

IL17f promoter **(-408 to -340)**

ACTGCATGACCCGAAAGCA	Chr1:20,774,671-20,774,688
AACCCACACGCAGAGCATGACAAGAG	*Chr1:20,774,643-20,774,669*
TTTAATTCCCCCACAAAGCAA	Chr1:20,774,620-20,774,640

IL17f -7 **(-7kb)**

TTCCCTTTTCTGCCTTGCA	Chr1:20,782,972-20,782,990
ACGAAGCACAGGGCTGGGCC	*Chr1:20,782,996-20,783,015*
TGTGTAACACGCAGAGTGGAATG	Chr1:20,783,017-20,783,039

IL21 promoter **(-529 to -382)**

GCATAGTCATCACCCCATAAA	Chr3:37,131,996-37,132,016
TCAGAGAAGTAAACACAAACAC	Chr3:37,131,869-37,131,890

IL22 promoter **(-600 to -417)**

GCACAGAATATAGGACACGGGT	Chr10:117,641,447-117,641,468
ACACAGTTTTCAAAGAAAGCCA	Chr10:117,641,609 117,641,630

Supplementary Table A2.S6. EMSA oligos.

IL-17 promoter oligos	Sequence 5' to 3'
33-1-top-IL17a	GCACCCAGCACCAGCTGATCAGGACGCG
33-1-bot-IL17a	GTTTGCGCGTCCTGATCAGCTGGTGCTG
46-14-top-IL17a	ACGAGGCACAAGTGCACCCAGCACCAGC
46-14-bot-IL17a	GATCAGCTGGTGCTGGGTGCACTTGTGC
69-37-top-IL17a	GCACTACTCTTCATCCACCTCACACGAG
69-37-bot-IL17a	TGTGCCTCGTGTGAGGTGGATGAAGAGT
83-51-top-IL17a	AAAGAGAGAAAGGAGCACTACTCTTCAT
83-51-bot-IL17a	GGTGGATGAAGAGTAGTGCTCCTTTCTC
100-68-top-IL17a	GTAGTAAAACCGTATAAAAAGAGAGAAA
100-68-bot-IL17a	GCTCCTTTCTCTCTTTTTATACGGTTTT
119-87-top-IL17a	ACGTAAGTGACCACAGAGGTAGTAAAA
119-87-bot-IL17a	TACGGTTTTACTACCTCTGTGGTCACT
140-106-top-IL17a	GTCACCCCCCAACCCACTCTTGACGTAAGT
140-106-bot-IL17a	TGGTCACTTACGTCAAGAGTGGGTTGGGGG
159-127-top-IL17a	GAATCTTTACTCAAATGGTGTCACCCCC
159-127-bot-IL17a	GGTTGGGGGGTGACACCATTTGAGTAAA
169-137-top-IL17a	TTTGAGGATGGAATCTTTACTCAAATGG
169-137-bot-IL17a	TGACACCATTTGAGTAAAGATTCCATCC
187-155-top-IL17a	GGTTCTGTGCTGACCTCATTTGAGGATG
187-155-bot-IL17a	GATTCCATCCTCAAATGAGGTCAGCACA
204-172-top-IL17a	GCCCGTCATAAAGGGGTGGTTCTGTGCT
204-172-bot-IL17a	AGGTCAGCACAGAACCACCCCTTTATGA
215-183-top-IL17a	AGACAGATGTTGCCCGTCATAAAGGGGT
215-183-bot-IL17a	GAACCACCCCTTTATGACGGGCAACATC
235-203-top-IL17a	GCCCTTCCCATCTACCTTCGAGACAGAT
235-203-bot-IL17a	GCAACATCTGTCTCGAAGGTAGATGGGA
250-217-top-IL17a	GCATAGTGAACTTCTGCCCTTCCCATCTA
250-217-bot-IL17a	GAAGGTAGATGGGAAGGGCAGAAGTTCAC
266-234-top-IL17a	GAAGTCATGCTTCTTTGCATAGTGAACT
266-234-bot-IL17a	GCAGAAGTTCACTATGCAAAGAAGCATG
281-249-top-IL17a	CTGTTCAGCTCCCAAGAAGTCATGCTTC
281-249-bot-IL17a	GCAAAGAAGCATGACTTCTTGGGAGCTG
302-269-top-IL17a	CTGAATCACAGCAAAGCATCTCTGTTCAG
302-269-bot-IL17a	GGGAGCTGAACAGAGATGCTTTGCTGTGA
320-286-top-IL17a	GTCCATACACACATGATACTGAATCACAGC
320-286-bot-IL17a	GCTTTGCTGTGATTCAGTATCATGTGTGTA
334-302-top-IL17a	GCAGCTTCAGATATGTCCATACACACAT
334-302-bot-IL17a	GTATCATGTGTGTATGGACATATCTGAA

349-317-top-IL17a	GAGCCCAGCTCTGCAGCAGCTTCAGATA
349-317-bot-IL17a	GGACATATCTGAAGCTGCTGCAGAGCTG
370-337-top-IL17a	GACTCACAAACCATTACTATGGAGCCCAG
370-337-bot-IL17a	CAGAGCTGGGCTCCATAGTAATGGTTTGT
383-351-top-IL17a	GAGACTGTCAAGAGACTCACAAACCATT
383-351-bot-IL17a	ATAGTAATGGTTTGTGAGTCTCTTGACA
400-368-top-IL17a	AAAGTGTGTGTCACTAGGAGACTGTCAA
400-368-bot-IL17a	GTCTCTTGACAGTCTCCTAGTGACACAC
416-384-top-IL17a	GATCAAGTCAAAATTCAAAGTGTGTGTC
416-384-bot-IL17a	CTAGTGACACACACTTTGAATTTTGACT
433-401-top-IL17a	GGTAGAAAAGTGAGAAAGATCAAGTCAA
433-401-bot-IL17a	GAATTTTGACTTGATCTTTCTCACTTTT
445-413-top-IL17a	GCCAGGGAATTTGGTAGAAAAGTGAGAA
445-413-bot-IL17a	GATCTTTCTCACTTTTCTACCAAATTCC
464-432-top-IL17a	GGGCAAGGGATGCTCTCTAGCCAGGGAA
464-432-bot-IL17a	GCAAATTCCCTGGCTAGAGAGCATCCCT
476-44-top-IL17a	GTGGGTTTCTTTGGGCAAGGGATGCTCT
476-44-bot-IL17a	GCTAGAGAGCATCCCTTGCCCAAAGAAA
497-465-top-IL17a	GTTTACATACTAAGACATTGAGTGGGTT
497-465-bot-IL17a	AAAGAAACCCACTCAATGTCTTAGTATG

IL-21 promoter oligos	
33-1-top-IL21	GTCATCAGCTCCTGGAGACTCAGTTCTG
33-1-bottom-IL21	GCCACCAGAACTGAGTCTCCAGGAGCTG
55-22-top-IL21	GTGAGAACCAGACCAAGGCCCTGTCATCA
55-22-bottom-IL21	GGAGCTGATGACAGGGCCTTGGTCTGGTT
67-35-top-IL21	AGTCAGGTTGAAGTGAGAACCAGACCAA
67-35-bottom-IL21	GGGCCTTGGTCTGGTTCTCACTTCAACC
88-56-top-IL21	TAGCGACAACCTGTGCACAGTCAGGT
88-56-bottom-IL21	GTTCAACCTGACTGTGCACAGGTTGT
105-73-top-IL21	GATGAATAAATAGGTAGCCGTAGCGACA
105-73-bottom-IL21	CAGGTTGTCGCTACGGCTACCTATTTAT
120-88-top-IL21	GGCCTCTTCTTGAGGGATGAATAAATAG
120-88-bottom-IL21	GCTACCTATTTATTCATCCCTCAAGAAG
137-105-top-IL21	CTGCAATGGGAGGGCTTGGCCTCTTCTT
137-105-bottom-IL21	GCCTCAAGAAGAGGCCAAGCCCTCCCAT
150-118-top-IL21	AAAGATTTCCAGGCTGCAATGGGAGGGC
150-118-bottom-IL21	GCCAAGCCCTCCCATTGCAGCCTGGAAA
174-142-top-IL21	GTTACTCACACTCATCCACTATACAAAG
174-142-bottom-IL21	GAAATCTTTGTATAGTGGATGAGTGTGA
183-151-top-IL21	GAAAAACGAGTTACTCACACTCATCCAC

183-151-bottom-IL21	GTATAGTGGATGAGTGTGAGTAACTCGT
207-175-top-IL21	CACGTACACCTAGCCAATGGAAAAGAAA
207-175-bottom-IL21	TCGTTTTTCTTTTCCATTGGCTAGGTGT
221-189-top-IL21	TGCCCCCACACGCACACGTACACCTAGC
221-189-bottom-IL21	CATTGGCTAGGTGTACGTGTGCGTGTGG
240-208-top-IL21	TGTGGACTCTATCCATCCCTGCCCCCAC
240-208-bottom-IL21	TGCGTGTGGGGGCAGGGATGGATAGAGT
254-222-top-IL21	GATGGGGCACATTTTGTGGACTCTATCC
254-222-bottom-IL21	GGGATGGATAGAGTCCACAAAATGTGCC
266-234-top-IL21	GTCTAAGATGCAGATGGGGCACATTTTG
266-234-bottom-IL21	GTCCACAAAATGTGCCCCATCTGCATCT
279-247-top-IL21	GTCTCTTTTTCCTGTCTAAGATGCAGAT
279-247-bottom-IL21	GCCCCATCTGCATCTTAGACAGGAAAAA
304-272-top-IL21	GCTGAAAACTGGAATTCACCCATGTGTC
304-272-bottom-IL21	AAAGAGACACATGGGTGAATTCCAGTTT
314-282-top-IL21	CTTGGTGAATGCTGAAAACTGGAATTCA
314-282-bottom-IL21	ATGGGTGAATTCCAGTTTTCAGCATTCA
334-303-top-IL21	GACACACACACACACACACACCTTGGTG
334-303-bottom-IL21	GCATTCACCAAGGTGTGTGTGTGTGTGTG
361-328-top-IL21	GCCACACACACACACACACACACACACA
361-328-bottom-IL21	GTGTGTGTGTGTGTGTGTGTGTGTGTGT
383-351-top-IL21	GAAATCTGACGGTGCCTCCTGTGCCACA
383-351-bottom-IL21	GTGTGTGTGGCACAGGAGGCACCGTCAG
395-363-top-IL21	GTTTACTTCTCTGAAATCTGACGGTGCC
395-363-bottom-IL21	CAGGAGGCACCGTCAGATTTCAGAGAAG
410-378-top-IL21	GATCAAAGTGTTTGTGTTTACTTCTCTG
410-378-bottom-IL21	GATTTCAGAGAAGTAAACACAAACACTT
422-390-top-IL21	TGCAGAGCAAAAGATCAAAGTGTTTGTG
422-390-bottom-IL21	GTAAACACAAACACTTTGATCTTTTGCT
447-415-top-IL21	GACAAACCAGGTGAGGTGCCAGGGATGC
447-415-bottom-IL21	GCTCTGCATCCCTGGCACCTCACCTGGT
463-429-top-IL21	GCCTTTATGACTGTCAGACAAACCAGGTGA
463-429-bottom-IL21	GCACCTCACCTGGTTTGTCTGACAGTCATA
476-445-top-IL21	GTCATTGCAGAAGTGCCTTTATGACTGT
476-445-bottom-IL21	GTCTGACAGTCATAAAGGCACTTCTGCA
494-462-top-IL21	GCCATGCCGCTGCTTTACTCATTGCAGA
494-462-bottom-IL21	GCACTTCTGCAATGAGTAAAGCAGCGGC
509-477-top-IL21	AAAGTTCCAATAAAGGCCATGCCGCTGC
509-477-bottom-IL21	GTAAAGCAGCGGCATGGCCTTTATTGGA
525-493-top-IL21	AGTCATCACCCCATAAAAAGTTCCAATA
525-493-bottom-IL21	GCCTTTATTGGAACTTTTTATGGGGTGA
543-511-top-IL21	GGTTCAGTCAAAAAGCATAGTCATCACC
543-511-bottom-IL21	TATGGGGTGATGACTATGCTTTTTGACT
558-526-top-IL21	AATGGAGTACAGGATGGTTCAGTCAAAA
558-526-bottom-IL21	ATGCTTTTTGACTGAACCATCCTGTACT
578-546-top-IL21	GTAACCTCTTCCATCATTGCAATGGAGT

578-546-bottom-IL21	CCTGTACTCCATTGCAATGATGGAAGAG
604-573-top-IL21	GCCCATCATTTAATTCTTCCTAAGAAG
604-573-bottom-IL21	GGTTACTTCTTAGGAAGAATTAAATGA
618-586-top-IL21	AGGTTAGAAAACTAGCCCATCATTTAAT
618-586-bottom-IL21	GAAGAATTAAATGATGGGCTAGTTTTCT
639-607-top-IL21	AGGATCTAAAATACTCTTGCTAGGTTAG
639-607-bottom-IL21	GTTTTCTAACCTAGCAAGAGTATTTTAG
657-625-top-IL21	GCACCCTTACAAAAAGATAAGGATCTAA
657-625-bottom-IL21	GTATTTTAGATCCTTATCTTTTTGTAAG
678-646-top-IL21	TGGAAGCAAATCCTATTTTAACACCCTT
678-646-bottom-IL21	TTTGTAAGGGTGTTAAAATAGGATTTGC
705-672-top-IL21	GCTATTTAAAGATACACTGGTGAAAATTG
705-672-bottom-IL21	GCTTCCAATTTTCACCAGTGTATCTTTAA
718-686-top-IL21	AGGCACCATTAGTGCTATTTAAAGATAC
718-686-bottom-IL21	CCAGTGTATCTTTAAATAGCACTAATGG
736-704-top-IL21	GTTACATAAAGTGTCAGGAGGCACCATT
736-704-bottom-IL21	GCACTAATGGTGCCTCCTGACACTTTAT
754-722-top-IL21	GTATTTACAATCCATATTGTTACATAAA
754-722-bottom-IL21	GACACTTTATGTAACAATATGGATTGTA
775-743-top-IL21	AGTTCATCAAAACTGTTTATTGTATTTA
775-743-bottom-IL21	GATTGTAAATACAATAAACAGTTTTGAT
792-760-top-IL21	GAGCACGCTGTCTACTTAGTTCATCAAA
792-760-bottom-IL21	ACAGTTTTGATGAACTAAGTAGACAGCG

IL-22 promoter oligos	
33-1-top-IL22	AGTTATCAACTGTTGACACTTGTGCGAT
33-1-bottom-IL22	CAGAGATCGCACAAGTGTCAACAGTTGA
48-16-top-IL22	ACAGGCTCTCCTCTCAGTTATCAACTGT
48-16-bottom-IL22	TGTCAACAGTTGATAACTGAGAGGAGAG
69-37-top-IL22	TTGCCTTTTGCTCTCTCACTAACAGGCT
69-37-bottom-IL22	AGGAGAGCCTGTTAGTGAGAGAGCAAAA
85-53-top-IL22	TGCTCCCCTGATGTTTTTGCCTTTTGCT
85-53-bottom-IL22	GAGAGAGCAAAAGGCAAAAACATCAGGG
107-75-top-IL22	GTACCATGCTACCCGACGAACATGCTCC
107-75-bottom-IL22	TCAGGGGAGCATGTTCGTCGGGTAGCAT
123-91-top-IL22	GACAATCATCTGCTTGGTACCATGCTAC
123-91-bottom-IL22	GTCGGGTAGCATGGTACCAAGCAGATGA
146-114-top-IL22	AGGTAAGCACTCAGACCTCTACAGACAA
146-114-bottom-IL22	GATGATTGTCTGTAGAGGTCTGAGTGCT
160-128-top-IL22	AGAGACACCTAAACAGGTAAGCACTCAG
160-128-bottom-IL22	GAGGTCTGAGTGCTTACCTGTTTAGGTG
181-149-top-IL22	TCTGCCTCTCCCATCACAAGCAGAGACA
181-149-bottom-IL22	TTAGGTGTCTCTGCTTGTGATGGGAGAG
193-161-top-IL22	AAAAGCAGCAACTTCTGCCTCTCCCATC
193-161-bottom-IL22	CTTGTGATGGGAGAGGCAGAAGTTGCTG
214-182-top-IL22	CCTGGTGTCCCGATGGCTATAAAAGCAG

214-182-bottom-IL22	AGTTGCTGCTTTTATAGCCATCGGGACA
233-201-top-IL22	GTCACAATACCAAAAAAACCCTGGTGTC
233-201-bottom-IL22	ATCGGGACACCAGGGTTTTTTTGGTATT
252-220-top-IL22	AATGTCTGATGTCATATCATTCACAATA
252-220-bottom-IL22	TTTGGTATTGTGAATGATATGACATCAG
267-235-top-IL22	GACTGGAAATTAGATAATGTCTGATGTC
267-235-bottom-IL22	GATATGACATCAGACATTATCTAATTTC
293-261-top-IL22	GTGGTTAGGTACTTCTCAGAAGACAGGA
293-261-bottom-IL22	TCCAGTCCTGTCTTCTGAGAAGTACCTA
305-273-top-IL22	TGGCCTCCTATGGTGGTTAGGTACTTCT
305-273-bottom-IL22	TTCTGAGAAGTACCTAACCACCATAGGA
329-297-top-IL22	GGAAGGCTTGGAGGTGGTGTCTTGTGGC
329-297-bottom-IL22	AGGAGGCCACAAGACACCACCTCCAAGC
340-309-top-IL22	GCTCTCAAGGTGGGAAGGCTTGGAGGTG
340-309-bottom-IL22	GACACCACCTCCAAGCCTTCCCACCTTG
366-334-top-IL22	GTGACGTTTTAGGGAAGACTTCCCATCT
366-334-bottom-IL22	TTGAGAGATGGGAAGTCTTCCCTAAAAC
380-348-top-IL22	TGTTGGCCCTCACCGTGACGTTTTAGGG
380-348-bottom-IL22	GTCTTCCCTAAAACGTCACGGTGAGGGC
405-373-top-IL22	CTGGGATTTGTGTGCAAAAGCACCTTGT
405-373-bottom-IL22	GGCCAACAAGGTGCTTTTGCACACAAAT
420-388-top-IL22	GTGTTTAGAAGATTTCTGGGATTTGTGT
420-388-bottom-IL22	TTTGCACACAAATCCCAGAAATCTTCTA
497-465-top-IL22	AATAGCTACGGGAGATCAAAGGCTGCTC
497-465-bottom-IL22	GAGTAGAGCAGCCTTTGATCTCCCGTAG
518-486-top-IL22	CCGTGACCAAAACGCTGACTCAATAGCT
518-486-bottom-IL22	CCCGTAGCTATTGAGTCAGCGTTTTGGT
528-495-top-IL22	GAAAATGAGTCCGTGACCAAAACGCTGAC
528-495-bottom-IL22	ATTGAGTCAGCGTTTTGGTCACGGACTCA
536-504-top-IL22	GTTGGTGGGAAAATGAGTCCGTGACCAA
536-504-bottom-IL22	GCGTTTTGGTCACGGACTCATTTTCCCA
540-506-top-IL22	TGAAGTTGGTGGGAAAATGAGTCCGTGACC
540-506-bottom-IL22	GTTTTGGTCACGGACTCATTTTCCCACCAA
547-513-top-IL22	GAATCTATGAAGTTGGTGGGAAAATGAGTC
547-513-bottom-IL22	TCACGGACTCATTTTCCCACCAACTTCATA
558-527-top-IL22	TAAAGAGATAAGAATCTATGAAGTTGGT
558-527-bottom-IL22	GTCCCACCAACTTCATAGATTCTTATCT
574-543-top-IL22	GTATTTCTGGTCACTTCTAAAGAGATAA
574-543-bottom-IL22	GATTCTTATCTCTTTAGAAGTGACCAGA
595-563-top-IL22	GAATATAGGACACGGGTCTTTTATTTCT
595-563-bottom-IL22	TGACCAGAAATAAAAGACCCGTGTCCTA
612-580-top-IL22	GCTTATTTCAAAGCACAGAATATAGGAC
612-580-bottom-IL22	CCCGTGTCCTATATTCTGTGCTTTGAAA
628-596-top-IL22	CCAAGTTTTCATTATGGCTTATTTCAAA
628-596-bottom-IL22	TGTGCTTTGAAATAAGCCATAATGAAAA
650-619-top-IL22	GATTTTAAAAATTGAAATAATCTCCAAG

650-619-bottom-IL22	GAAAACTTGGAGATTATTTCAATTTTTA
662-630-top-IL22	AGAGATATAATTATTTTAAAAATTGAAA
662-630-bottom-IL22	GATTATTTCAATTTTTAAAATAATTATA
684-652-top-IL22	GGATTCCATATACTAAAAAAATAGAGATA
684-652-bottom-IL22	GATTATATCTCTATTTTTTTAGTATATGG
700-668-top-IL22	AGCTAGTTATAGTTTAGGATTCCATATA
700-668-bottom-IL22	TTTAGTATATGGAATCCTAAACTATAAC

AP-1 Consensus Probe[36]	
Top	AGCTTCGCTTGATGAGTC
Bottom	GCCGACTGAGTAGTTCGC

RORE element[38]	
Top	GAAAGTTTTCTGACCCACTTTAAATCA
Bottom	CTTTAACTAAATTTCACCCAGTCTTTT

Supplementary Figure Legends

Supplementary Figure A2.S1. Expression and cellular location of Batf in T cells.

a, The expression profile of *Batf* among the indicated tissues was determined by Affymetrix gene microarray. The data are presented in arbitrary units and reflect normalized and modeled expression values generated using DNA-Chip analyzer (dChip) software. **b, c,** Batf is located in the cytoplasm and nucleus of resting T cells. **b,** DO11.10 $CD4^+$ T cells from CD2-N-FLAG-*Batf* transgenic or littermate control mice were isolated and differentiated with OVA and APCs under TH2 conditions. On day 7 cells were either left untreated or stimulated with PMA/ionomycin for 4h. Cells were then allowed to settle on to poly-L-lysine treated slides and stained with an anti-FLAG antibody, anti-CD4 antibody and DAPI as a nuclear stain as described in Methods. **c,** Naïve DO11.10 $CD4^+$ T cells from CD2-N-FLAG-*Batf* transgenic or littermate control mice were isolated and stained as in **c**. Data are representative of 2 independent experiments.

Supplementary Figure A2.S2. Targeting of the *Batf* locus by homologous recombination.

a, The endogenous genomic *Batf* locus, targeting construct and the mutant allele before and after cre-mediated deletion of the neomycin cassette are shown. Restriction enzyme digestion with BamHI of the genomic locus results in a 14.3kb wild type fragment that is detected by Southern Blot probes A and B; in the targeted allele, probe A detects a 2kb and probe B detects a 9kb fragment. In the neomycin-deleted targeted allele, BamHI digestion results in a 9kb fragment that is detected by both the 5' and 3' Southern Blot

probes. The neomycin resistance cassette was deleted by in vitro treatment with a cre-expressing Adenovirus. **b,** Southern Blot analysis of targeted *Batf* alleles. Probe A was used to hybridize BamHI digested genomic DNA from the indicated genotypes resulting from *Batf*$^{+/-}$ intercrosses. **c,** No residual protein expression in *Batf*$^{-/-}$ mice. Equal cell numbers from total splenocytes activated under TH17 conditions for 3 days were lysed in RIPA buffer and analyzed by Western Blot using anti-Batf antibody. The blots were stripped and reblotted with an antibody to β-actin to show equal protein loading. **d,** Batf expression in naïve T cells. Magnetically purified *Batf*$^{+/+}$ and *Batf*$^{-/-}$ CD4^{+} T cells were lysed in RIPA buffer. 1.5x10^{6} cell equivalents were subjected to Western Blot analysis. Blots were stripped and reprobed with anti-β-actin to show equal protein loading. **e,** CD4^{+} T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were stimulated with anti-CD3/CD28 under TH2 conditions for 4 days, left untreated or restimulated with PMA/ionomycin for 4h. Nuclear extracts from 0.5x10^{6} cell equivalents were analyzed for Batf expression by Western Blot. The blots were stripped and reprobed with anti-Lamin B antibody to show equal protein loading. Data are representative of 2 independent experiments.

Supplementary Figure A2.S3. Thymus, spleen and lymph nodes develop normally in *Batf*$^{-/-}$ mice.

a, Total cell numbers of thymus (n=11) and spleen (n=17) from individual 8-10 week old *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice are shown (horizontal bars indicate mean cell numbers). **b,** *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were injected with Evans Blue dye solution into each hind foot pad. After 1.5 hrs, mice were sacrificed and superficial inguinal lymph nodes were visualized using a dissecting microscope[35]. Data are representative of 2 independent experiments.

Supplementary Figure A2.S4. T and B cell development is normal in *Batf*$^{-/-}$ mice.

a, Thymus, spleen and lymph nodes of mice of the indicated genotypes were analyzed for the surface expression of CD4 and CD8 by flow cytometry. The percentages of $CD8^+$, $CD4^+$ and $CD4^+CD8^+$ T cells were similar between *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice. **b,** Splenic $CD4^+$ and $CD8^+$ cells were analyzed for the surface expression of the activation markers CD62L (left panel) and CD44 (right panel) on *Batf*$^{+/+}$ and *Batf*$^{-/-}$cells. A histogram overlay of surface expression of CD62L and CD44 on *Batf*$^{+/+}$ and *Batf*$^{-/-}$$CD4^+$ and $CD8^+$ T cells is shown. **c,** Total splenocytes were stained for CD3 in conjunction with unloaded or PBS57-loaded CD1d tetramers. NKT cells are identified as $CD3^+$CD1d-$PBS57^+$. **d,** Total splenocytes were analyzed by staining with antibodies to B220, AA4.1, IgM and IgD. The percentages of immature B cells ($AA4.1^+$ $B220^+$), Transitional 1 ($B220^+IgM^{hi}IgD^{lo}$), Transitional 2 ($B220^+IgM^{hi}$, IgD^{hi}) or mature B cells ($AA4.1^-B220^+$; $B220^+IgM^{lo}IgD^{hi}$) were similar between *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice. **e,** Bone marrow cells were stained for the expression of B220, CD43 and either BP1 and CD24 or IgD and IgM. The percentages of cells included in $B220^+CD43^{hi}$ subsets: BP-1^-CD24^- (Hardy fraction A), BP-1^-CD24^+ (Hardy fraction B), and BP-1^+CD24^+ (Hardy fraction C) were similar between *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice. Also the percentages of $B220^+$ $CD43^-$ subsets; IgM^-IgD^-(Hardy fraction D), IgM^+IgD^{lo} (Hardy fraction E), and $IgM^{lo}IgD^{hi}$ (Hardy fraction F) were similar between *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice. Numbers of all FACS plots indicate percentage of cells in the indicated region or gate. Data are representative of at least 2 independent experiments performed with multiple mice of each genotype.

Supplementary Figure A2.S5. The development of myeloid cells is grossly normal in *Batf*$^{-/-}$ mice.

a, Conventional splenic dendritic cell (cDC) subsets are present at normal ratios in *Batf*$^{-/-}$ mice. Single cell suspensions from collagenase and DNase treated spleens were stained with the indicated antibodies. cDCs were identified as CD11c^{hi} cells and further subdivided into CD4^{+} DCs and CD8^{+} DCs, identified as CD11c^{hi}CD4^{+}CD8^{-}and CD11c^{hi}CD4^{-}CD8α^{+} respectively. CD8^{+} DCs were further identified as CD11c^{hi}CD8α^{+} Dec205^{+}. **b,** Splenic single cell suspensions were prepared as in **a** and stained with antibodies to CD11c, CD11b, Gr1 and B220. Percentages of plasmacytoid dendritic cells, identified as CD11b^{-}CD11c^{lo}B220^{+}Gr1^{+}, were similar between *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice. Numbers for all FACS plots indicate the percentage of live cells in each gate or region. Data are representative of at least 2 independent experiments performed with multiple mice of each genotype.

Supplementary Figure A2.S6. Batf regulates IL-17 production by CD4^{+} and CD8^{+} cells.

a, Naïve CD4^{+}CD62L^{+}CD25^{-}T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice activated with anti-CD3 and anti-CD28 alone or under TH1 or TH2 conditions. Cells were restimulated on day 7 with anti-CD3/CD28 for 24h and analyzed for IFN-γ and IL-4 production. **b,** CD4^{+} T cells from DO11.10 *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were purified by magnetic bead separation and activated with OVA and irradiated APCs under TH17 conditions. 3 days later, cells were split and allowed to expand for 4 days in the presence of TH17 inducing cytokines.

After 3 rounds of differentiation, cells were restimulated with PMA/ionomycin for 4 hours and analyzed for IFN-γ and IL-17 expression by flow cytometry. **c,** Total splenocytes from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were stimulated under TH17 conditions for 3 days. Cells were restimulated with PMA/ionomycin and analyzed for IL-17 and IFN-γ expression by intracellular cytokine staining and flow cytometry. Plots are gated on CD8^{+} cells. **d,** DO11.10 transgenic CD4^{+} T cells from CD2-N-FLAG-*Batf* transgenic (TG) or transgene-negative (WT) control mice were stimulated with OVA and APC under TH17 conditions. 3 days later, cells were restimulated with PMA/ionomycin and cytokine production was analyzed by flow cytometry as described in Methods. **e,** Total splenocytes from CD2-N-FLAG-*Batf* transgenic (TG) or transgene-negative (WT) control mice were stimulated and analyzed as in **c**. **f,** Small intestinal lamina propria cells were isolated from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice and stimulated with PMA/ionomycin and stained for IL-17 and IFN-γ production. Plots are gated on CD4^{+} lymphocytes. Numbers for all FACS plots indicate the percentage of live cells in each indicated gate. Data are representative of at least 2 independent experiments performed with multiple mice of each genotype.

Supplementary Figure A2.S7. *Batf*$^{-/-}$ mice are resistant to EAE.

a, Total splenocytes were isolated from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice 10 days after EAE induction, stimulated with PMA/ionomycin for 3 hours and analyzed for IL-17 and IFN-γ expression by intracellular cytokine staining. Plots are gated on CD4^{+} cells. **b,** Spleens were isolated from unimmunized *Batf*$^{+/+}$ and *Batf*$^{-/-}$ or from mice 10 days after EAE induction. Total splenocytes were stained for the expression of CD4 and Foxp3 and

analyzed by flow cytometry. **c,** Spleens were isolated from unimmunized *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice or from mice 40 days after EAE induction. The abundance of Foxp3^{+} cells is depicted as the ratio of CD4^{+}Foxp3^{+} cells in the total CD4^{+} T cell compartment. **d,** 4 days prior to EAE induction, *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice received either control buffer (PBS) or 1x10^{7} *Batf*$^{+/+}$ CD4^{+} T cells. 40 days after EAE induction splenic and CNS infiltrating lymphocytes were stimulated with PMA/ionomycin for 4h and analyzed for IL-17 and IFN-γ production. Genotypes and whether mice received PBS or CD4^{+} T cells are indicated, disease scores are given in parentheses. FACS plots in **a** and **d** are gated on CD4^{+} cells. FACS plots are representative of 2-3 mice analyzed per group. Numbers for FACS plots indicate percentage of cells in each indicated gate.

Supplementary Figure A2.S8. Proximal IL-6 receptor signaling is normal in *Batf*$^{-/-}$ T cells.

a, Splenocytes from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were stained with antibodies to CD4 and IL-6 receptor (IL-6R). A histogram overlay of IL-6R expression on CD4^{+} cells of the indicated genotypes is shown. **b,** Magnetically purified *Batf*$^{+/+}$ and *Batf*$^{-/-}$CD4^{+} T cells (left) and CD8^{+} T cells (right) were stimulated with anti-CD3/CD28 in the presence of IL-6 for the indicated times and stained with an antibody to phospho-STAT3 (black lines) by intracellular staining as described in Methods. Untreated cells (grey lines) served as a negative control. **c,** Magnetically purified *Batf*$^{+/+}$ and *Batf*$^{-/-}$CD4^{+} T cells were stimulated with anti-CD3/CD28 in the presence of IL-21 for the indicated times and stained with an antibody to phospho-STAT3 (black lines) by intracellular staining. Untreated cells (grey lines) served as a negative control. **d,** Naïve CD4^{+}CD62L^{+}CD25^{-}T cells from *Batf*$^{+/+}$ and

Batf$^{-/-}$ mice were stimulated with TGF-β or TGF-β plus IL-6 for 3 days. Cells were stained for Foxp3 and analyzed by flow cytometry. **e,** Naïve CD4^{+}CD62L^{+}CD25^{-}T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were stimulated with TGF-β plus IL-6 in the presence of a neutralizing antibody to IL-2 for 3 days. Cells were stained for Foxp3, IL-17 and IFN-γ and analyzed by flow cytometry. Numbers for all FACS plots indicate the percentage of live cells in each indicated gate. Data are representative of at least 2 independent experiments performed with multiple mice of each genotype.

Supplementary Figure A2.S9. Batf does not regulate expression of genes induced by TGF-β alone or regulate SOCS gene expression.

a, b, Gene expression microarray analysis of T cells activated with anti-CD3/CD28 for 72h in the presence of the indicated cytokines and antibodies. **a,** A representative heat map of genes at least 5-fold induced by TGF-β compared to neutral conditions in *Batf*$^{+/+}$ T cells is presented. **b,** A representative heat map of the expression of suppressor of cytokine signaling (SOCS) genes in *Batf*$^{+/+}$ and *Batf*$^{-/-}$T cells is presented. **c,** Relative expression of RORγt, RORα and IL-22 in T cells 72h after activation with anti-CD3/CD28 under TH17 conditions was assessed by qRT-PCR. Data are normalized to HPRT and presented as percent expression relative to *Batf*$^{+/+}$ cells (mean + s.d. of 3 individual mice).

Supplementary Figure A2.S10. Several aspects of the IL-6-induced liver acute phase response are normal in *Batf*$^{-/-}$mice.

a, *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were injected intraperitonally with either 0.3ug IL-6 or saline.

4h after injection of mice the expression of the indicated acute phase proteins in liver was assessed by quantitative real time PCR. The relative expression of proteins normalized to HPRT is presented in arbitrary units. **b,** Relative expression of Batf in liver 4h after injection of mice with 0.3ug IL-6 or saline. The relative expression of proteins normalized to HPRT is presented in arbitrary units. Data represent mean + s.d. of 3 individual mice from independent experiments.

Supplementary Figure A2.S11. Retroviral overexpression of RORγt only partially restores IL-17 production in *Batf*$^{-/-}$T cells.

a, Naïve CD4^{+}CD62L^{+}CD25^{-}T cells were stimulated with anti-CD3/CD28 under TH17 conditions for 0, 8, 16, 24 and 62h. Relative expression (normalized to HPRT) of RORγt in *Batf*$^{+/+}$ and *Batf*$^{-/-}$T cells is depicted (error bars: mean ± s.d. of 3 individual mice). **b,** Magnetically purified CD4^{+} T cells were stimulated with anti-CD3/CD28 under TH17 conditions and were either left untreated or infected with empty-IRES-GFP-retrovirus (GFP-RV) or RORγt expressing IRES-GFP-retrovirus (RORγt-RV) as described in Methods. Cells were restimulated with PMA/ionomycin for 4h and analyzed for cytokine expression on day 3. **c,** CD4^{+} T cells were stimulated as indicated and infected with retrovirus as in **(b)** and Fig. 3e. The percentage of IL-17 producing cells among stably infected (GFP^{+}) cells is shown (mean + s.d. of three independent experiments). **d,** Dual retroviral overexpression of Batf and RORγt in *Batf*$^{-/-}$T cells. Magnetically purified *Batf*$^{-/-}$CD4^{+} T cells were stimulated with anti-CD3/CD28 under TH17 conditions and either infected with Batf-expressing IRES-GFP-retrovirus (Batf-RV), RORγt-expressing IRES-hCD4-retrovirus (RORγt-RV) or both retroviruses (bottom panel) as described in

Methods. As a control, cells were infected with empty-control retroviruses as indicated (top panel). Cells were restimulated with PMA/ionomycin and analyzed for IL-17 and IFN-γ expression on day 3. Data are representative of 2 independent experiments. Representative FACS plots shown are gated as indicated. Numbers represent percentage of cells in each gate or quadrant.

Supplementary Figure A2.S12. Batf binds several conserved non-coding regions in the IL-17 locus.

a, Vista blot depicting the sequence conservation of the human and mouse IL-17 loci. The locations of primers used for ChIP analysis are indicated. **b,** Magnetically purified $CD4^+$ T cells from *Batf*$^{+/+}$ or *Batf*$^{-/-}$ mice were activated with anti-CD3/CD28 coated beads under TH17 conditions (IL-6/TGF-β) for 24h, then subjected to ChIP analysis using anti-Batf polyclonal antibody as in Fig. 4b. Data are presented as relative binding based on normalization to unprecipitated input DNA (mean + s.d. of 2 independent experiments). **c,** *Batf*$^{+/+}$ $CD4^+$ T cells from C57Bl/6 mice were stimulated with anti-CD3 and APCs under TH17 conditions for 5 days. ChIP analysis of T cells before and after PMA/ionomycin stimulation for 4h was performed using anti-*Batf* antibody. The analyzed sites are denoted relative to the ATG for the *Il17a* or *Il17f* genes. Data are presented as relative binding based on normalization to unprecipitated input DNA (mean + s.d. of 2 independent experiments).

Supplementary Figure A2.S13. Identification of potential Batf binding sites in the IL-17a, IL-21 and IL-22 promoters.

a, $CD4^+$ T cells from DO11.10 *Batf*$^{+/+}$ and *Batf*$^{-/-}$ littermates were purified by magnetic

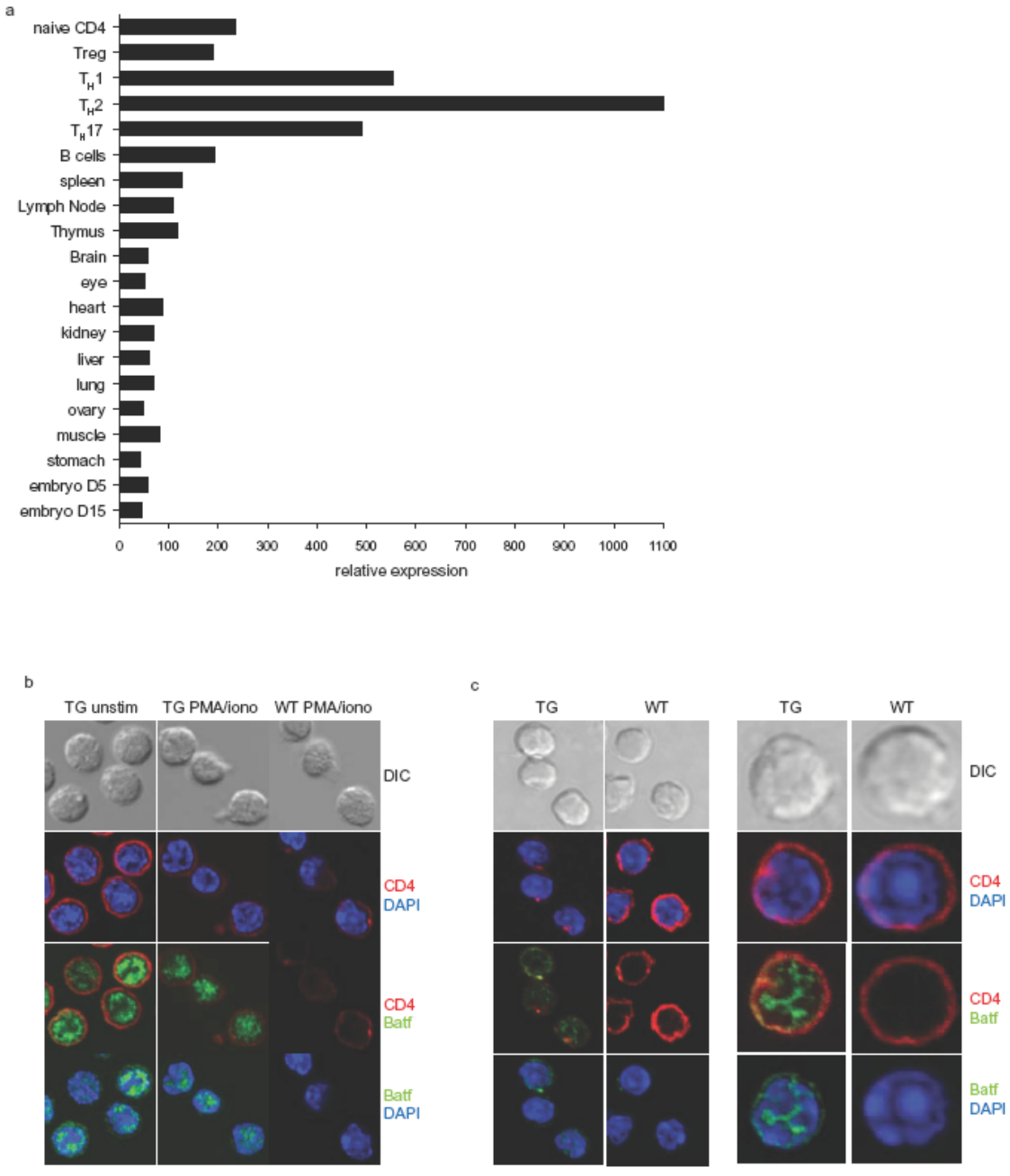
bead separation and activated with OVA and irradiated APCs under TH17 conditions. 3 days later, cells were split and allowed to expand for 4 days in the presence of TH17 inducing cytokines. On day 7 after initial stimulation cells were either left untreated or restimulated with PMA/ionomycin for 4 hours. Total cell extracts were analyzed for DNA binding ability to a consensus AP-1 probe (AGCTTCGCTTGATGAGTCAGCCG)[36] by electrophoretic mobility shift assay. **b-d,** Identification of potential Batf binding sites in the IL-17, IL-21 and IL-22 promoters. Total splenocytes from *Batf*-transgenic mice were stimulated under TH17 conditions for 3 days. Total cell extracts were analyzed for DNA binding ability to a consensus AP-1 probe[36] by EMSA as in **a**. Batf containing complexes were identified by supershift with anti-FLAG antibody. Sequences from the IL-17a **(b)**, IL-21 **(c)** and IL-22 **(d)** promoters were used to assess their ability to inhibit formation of Batf containing complexes as described in Methods. Sequences of competitors used are supplied in Supplementary Table 6. **e,** *Batf*$^{+/+}$ and *Batf*$^{-/-}$CD4^{+} T cells were stimulated under TH17 conditions for 5 days. ChIP analysis was performed as above. The analyzed sites are denoted relative to the ATG for the *IL21* or *IL22* genes. Data are presented as relative binding based on normalization to unprecipitated input DNA. **f**, WebLogo[37] presentation of the Batf-binding motif in the IL-17, IL-21 and IL-22 promoters. The size of each nucleotide is proportional to the frequency of its appearance at each position.

Supplementary Figure A2.S14. *Batf*$^{-/-}$ T cell do not protect against EAE.

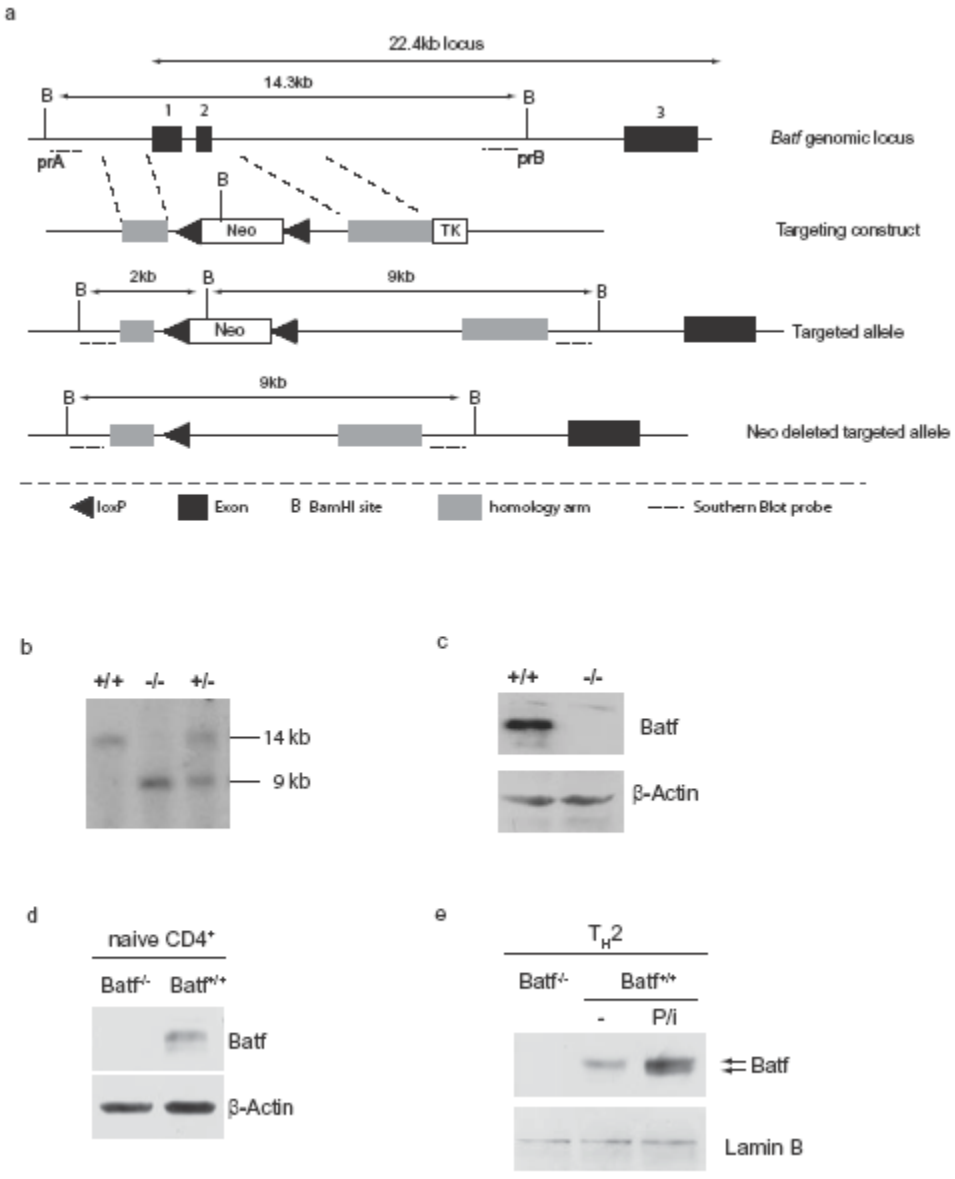
a, *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were injected 1x10^{7} *Batf*$^{-/-}$CD4^{+} T cells (n=4). Four days later, mice were immunized MOG33-35 peptide as described in Methods. Clinical EAE scores (mean +/- s.e.m) representative of two independent experiments are shown. **b,**

Magnetically purified $CD4^+$ T cells from wild type (WT) or mice lacking *Batf* (KO) were stimulated with anti-CD3/CD28 under TH17 conditions and either infected with control virus (GFP-RV) or Batf-expressing IRES-GFP-retrovirus (Batf-RV) as described in Methods. On day 7, cells re-stimulated with anti-CD3/CD28 under TH17 conditions. After 4 days, cells were restimulated with PMA/ionomycin and analyzed for CD4 and IL-17 expression. Representative FACS plots are gated on GFP^+ cells (KO) or GFP^- cells (wild type uninfected). Numbers represent percentage of cells in each region.

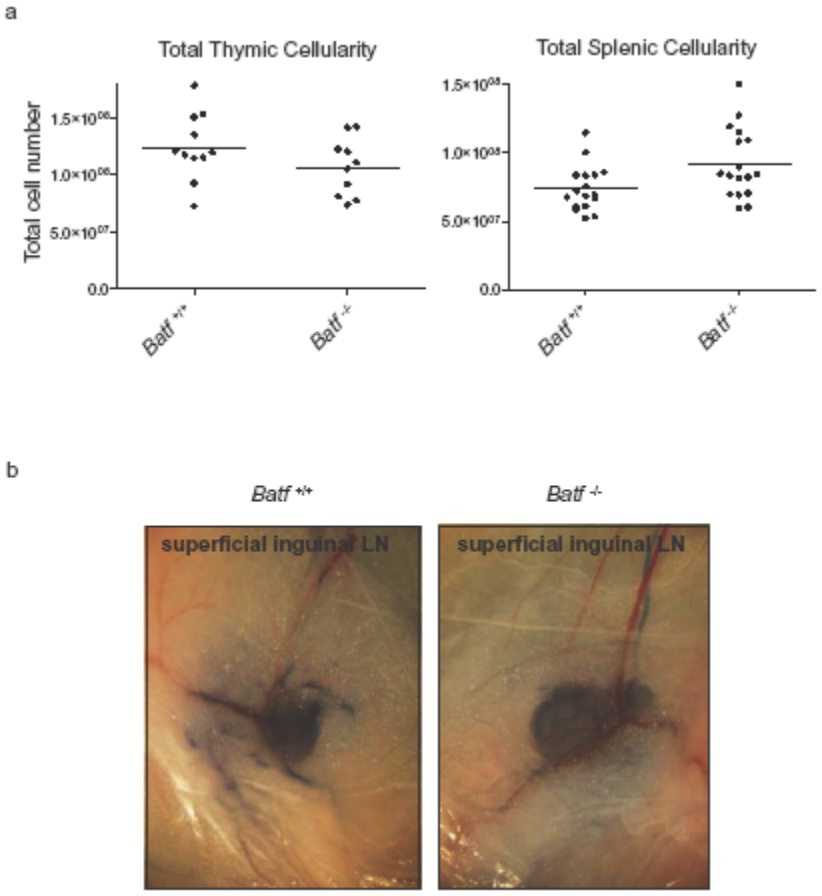
Supplementary Figure A2.S1.



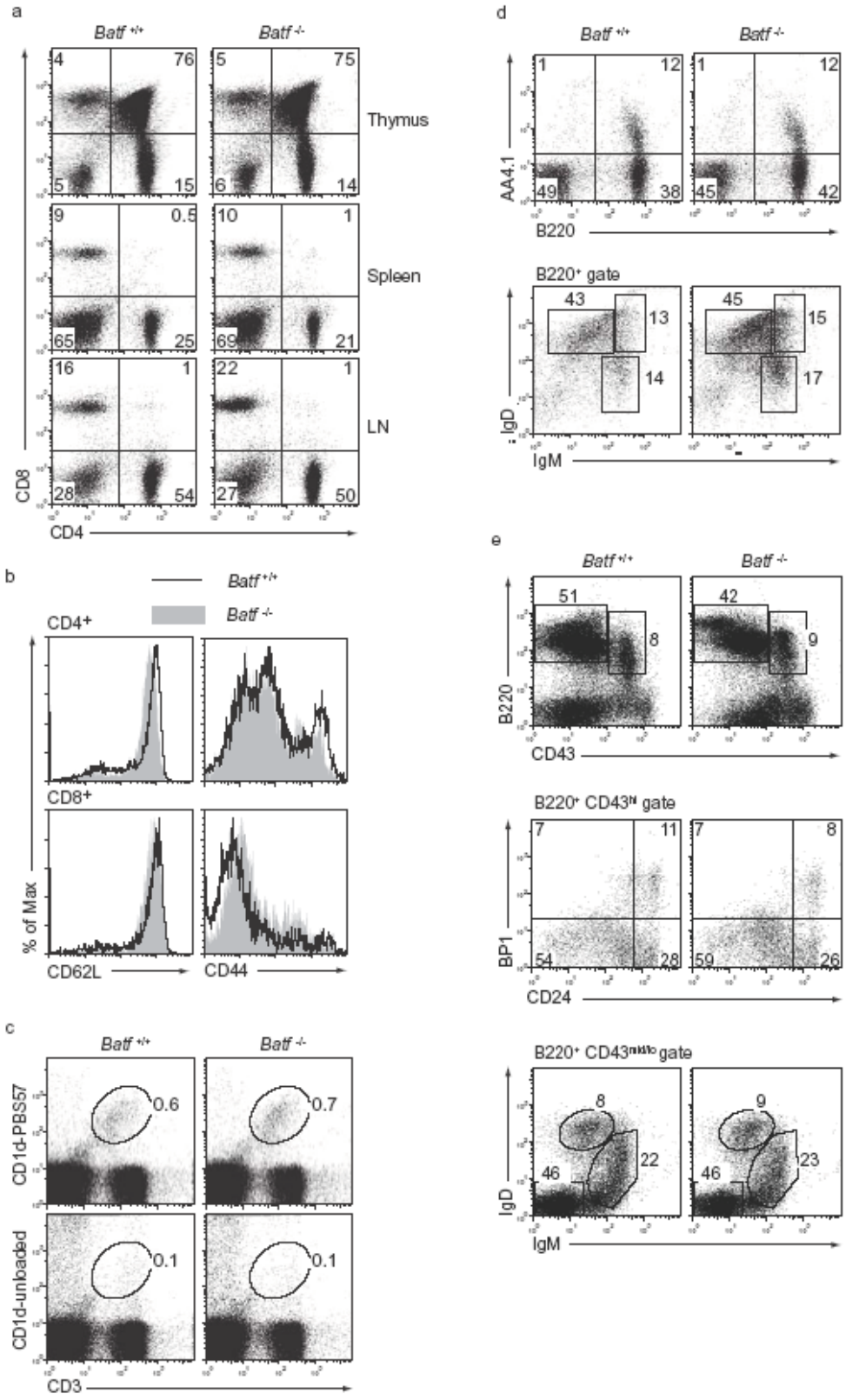
Supplementary Figure A2.S2.



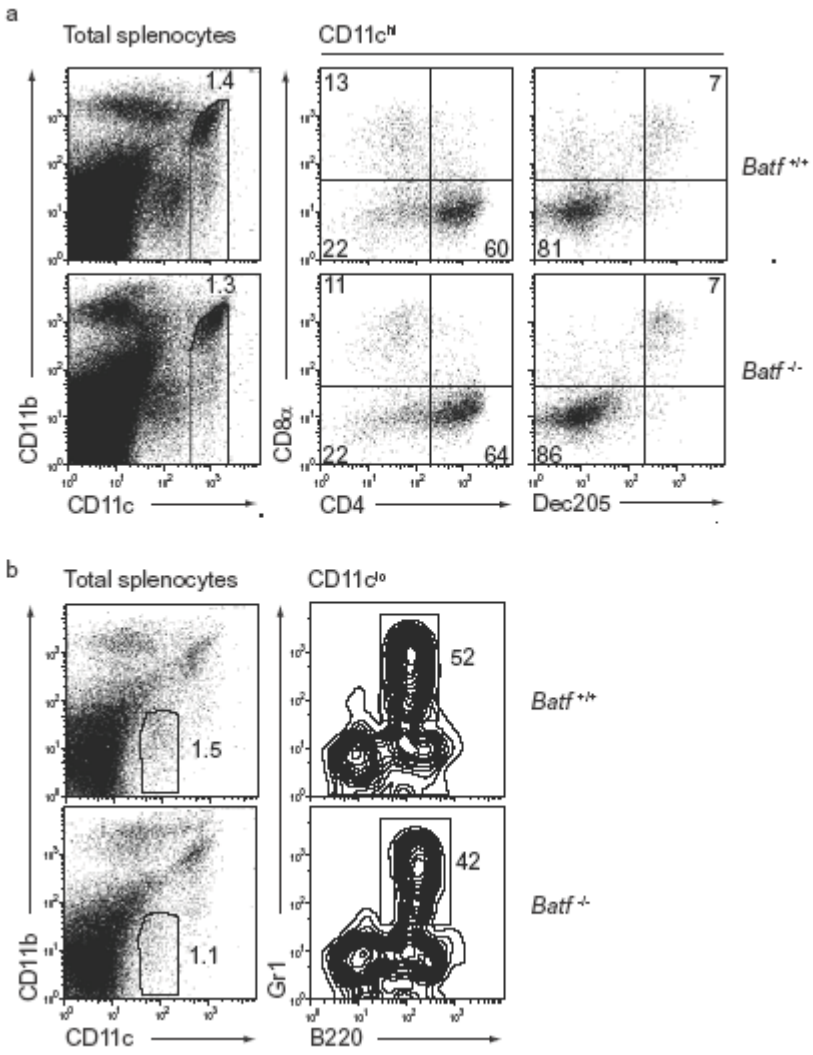
Supplementary Figure A2.S3.



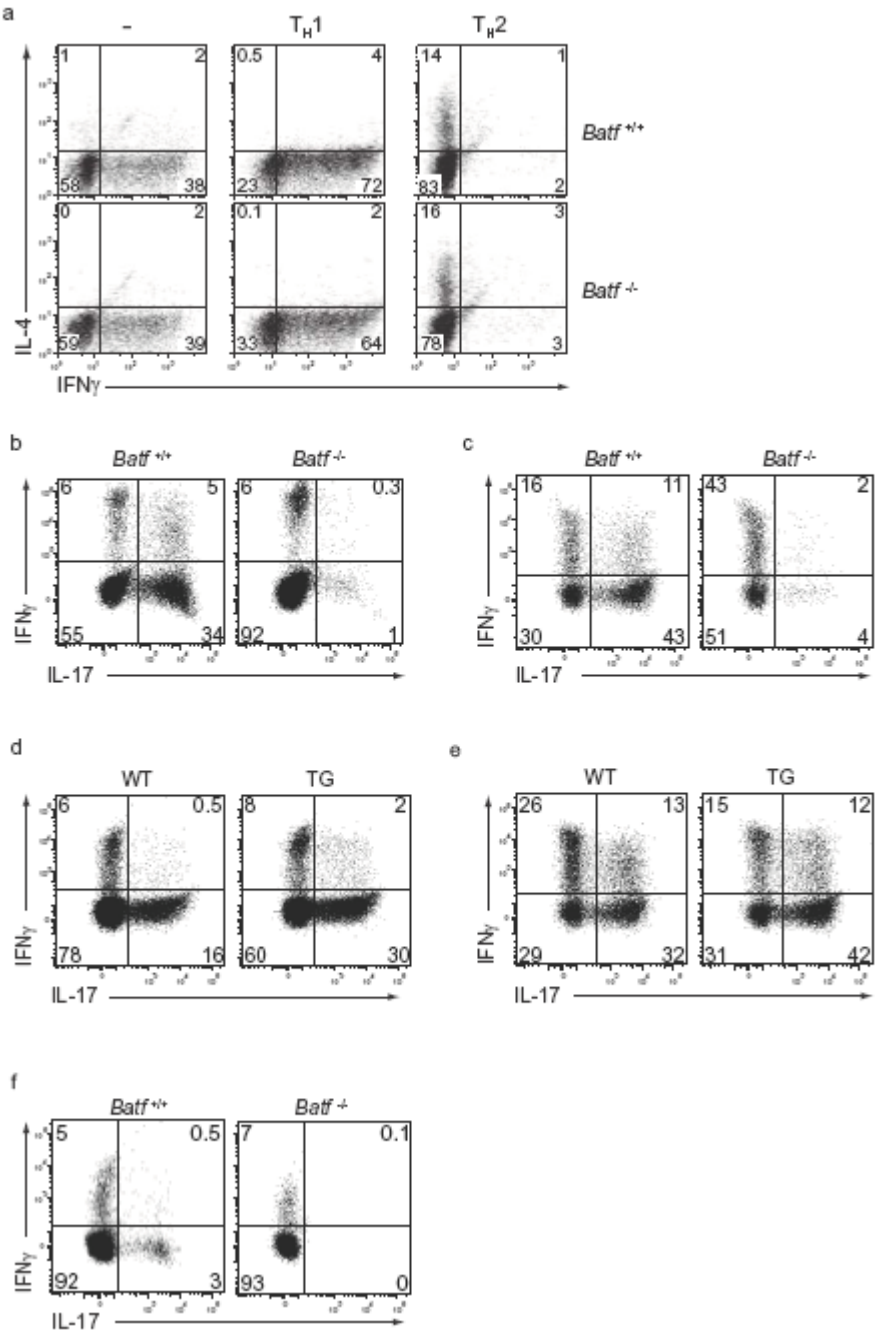
Supplementary Figures A2.S4.



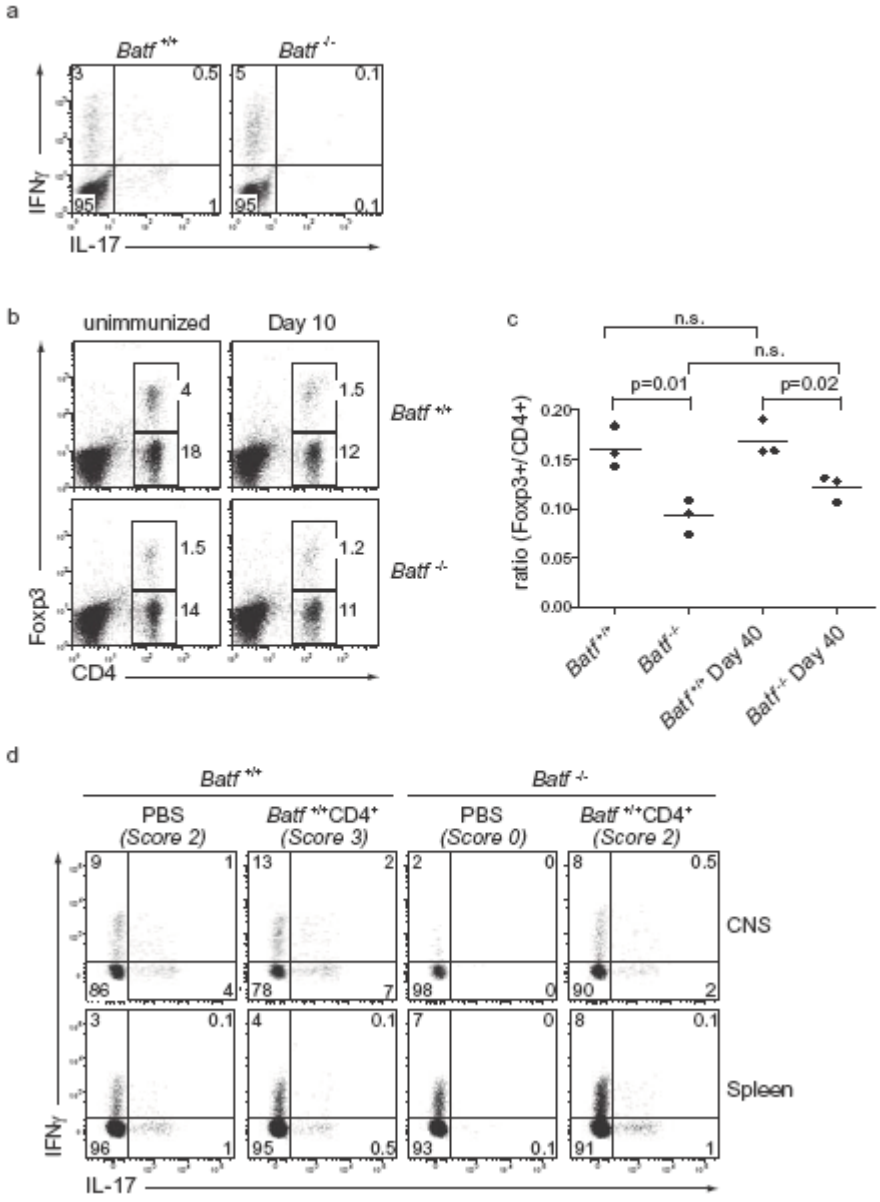
Supplementary Figure A2.S5.



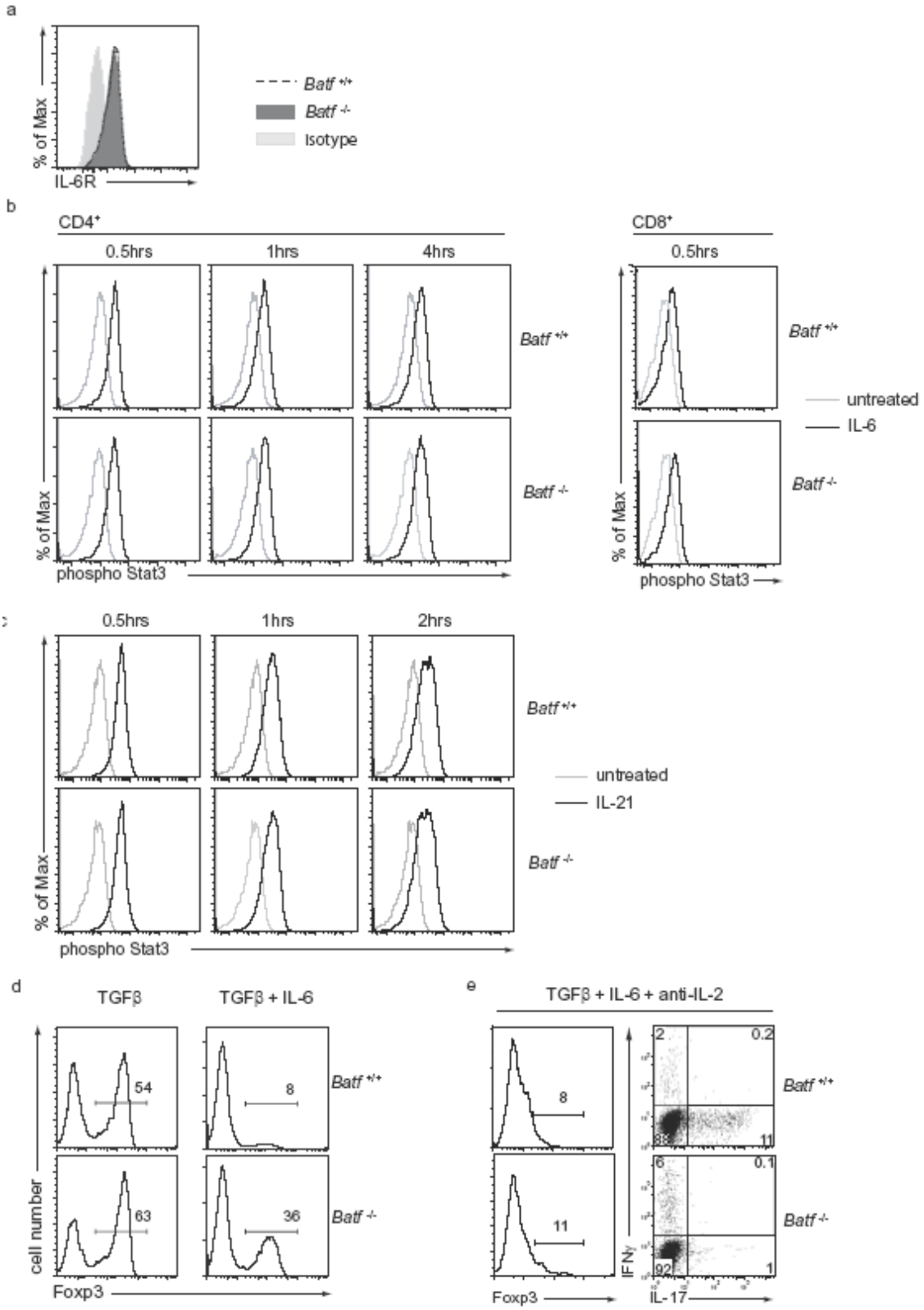
Supplementary Figure A2.S6.



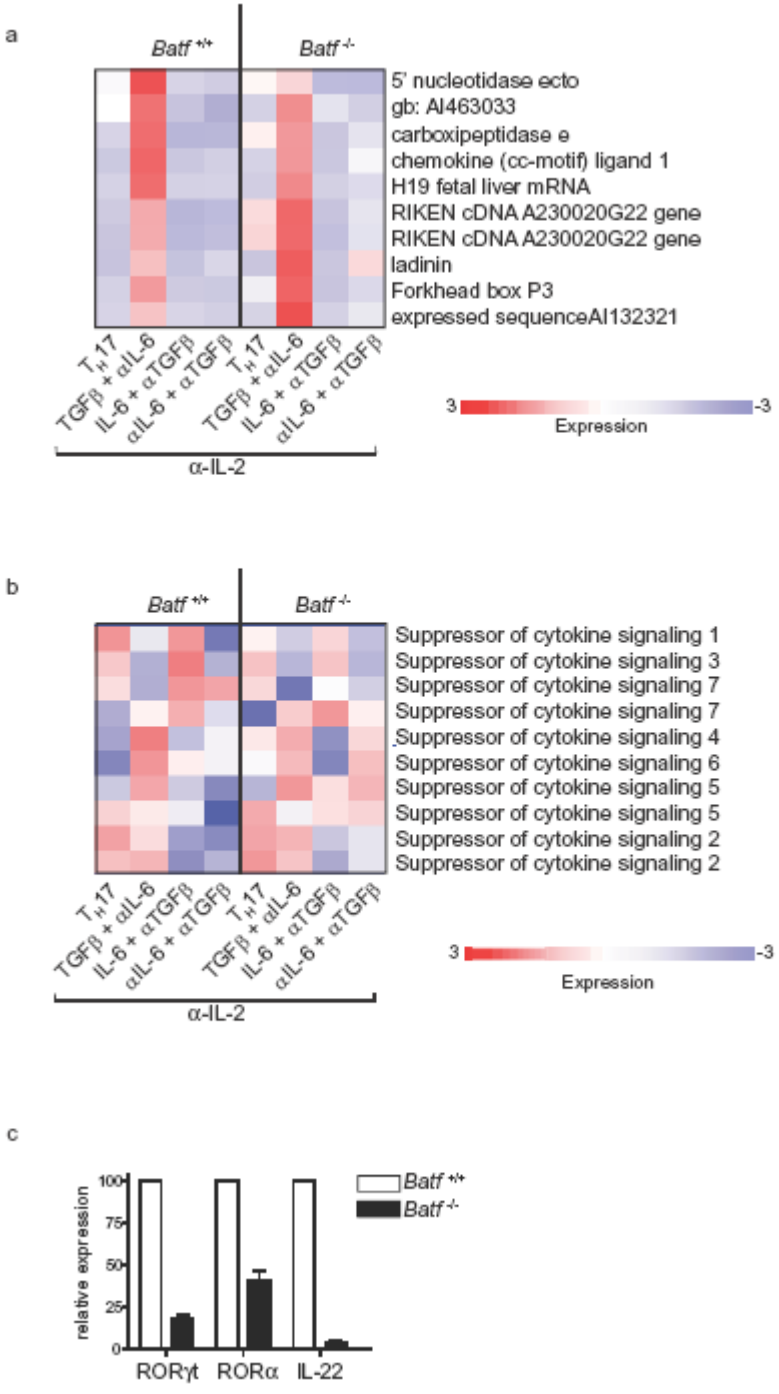
Supplementary Figure A2.S7.



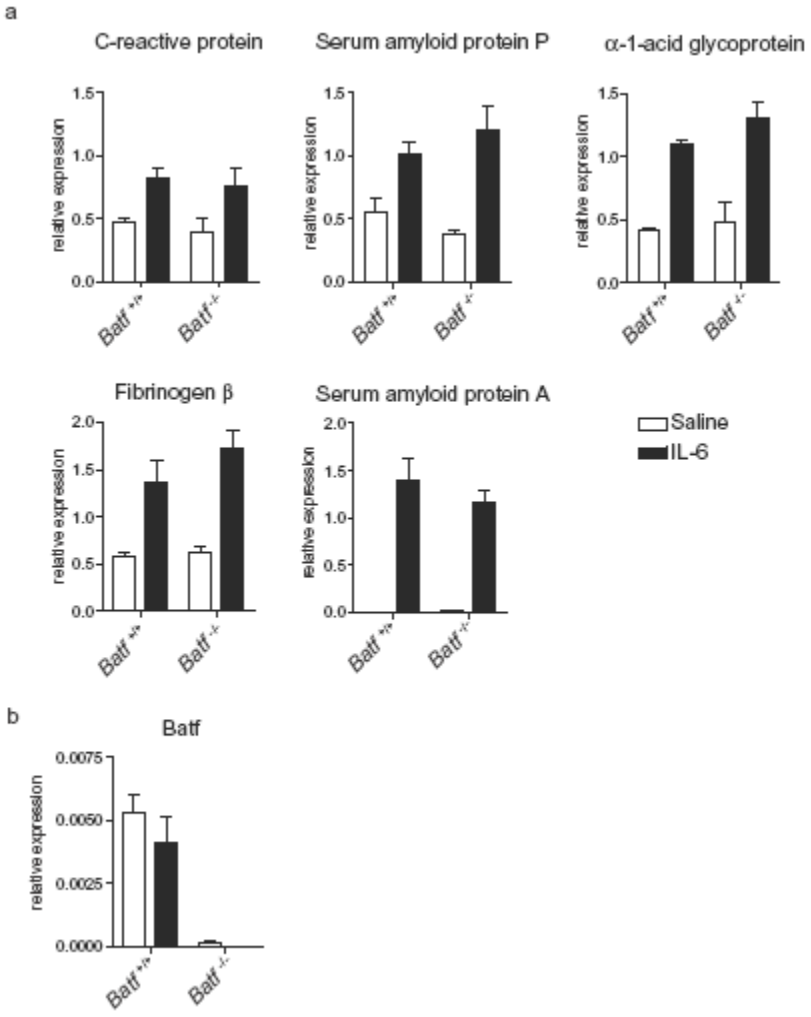
Supplementary Figure A2.S8.



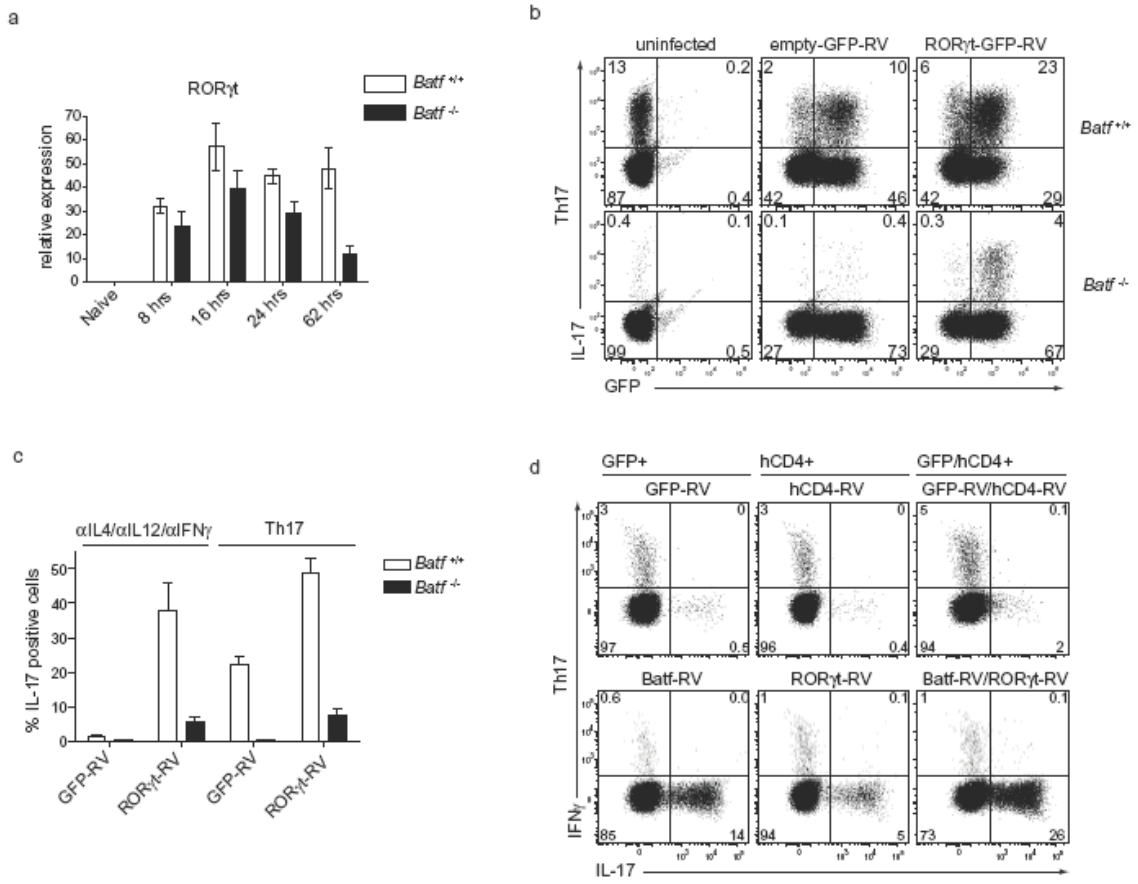
Supplementary Figure A2.S9.



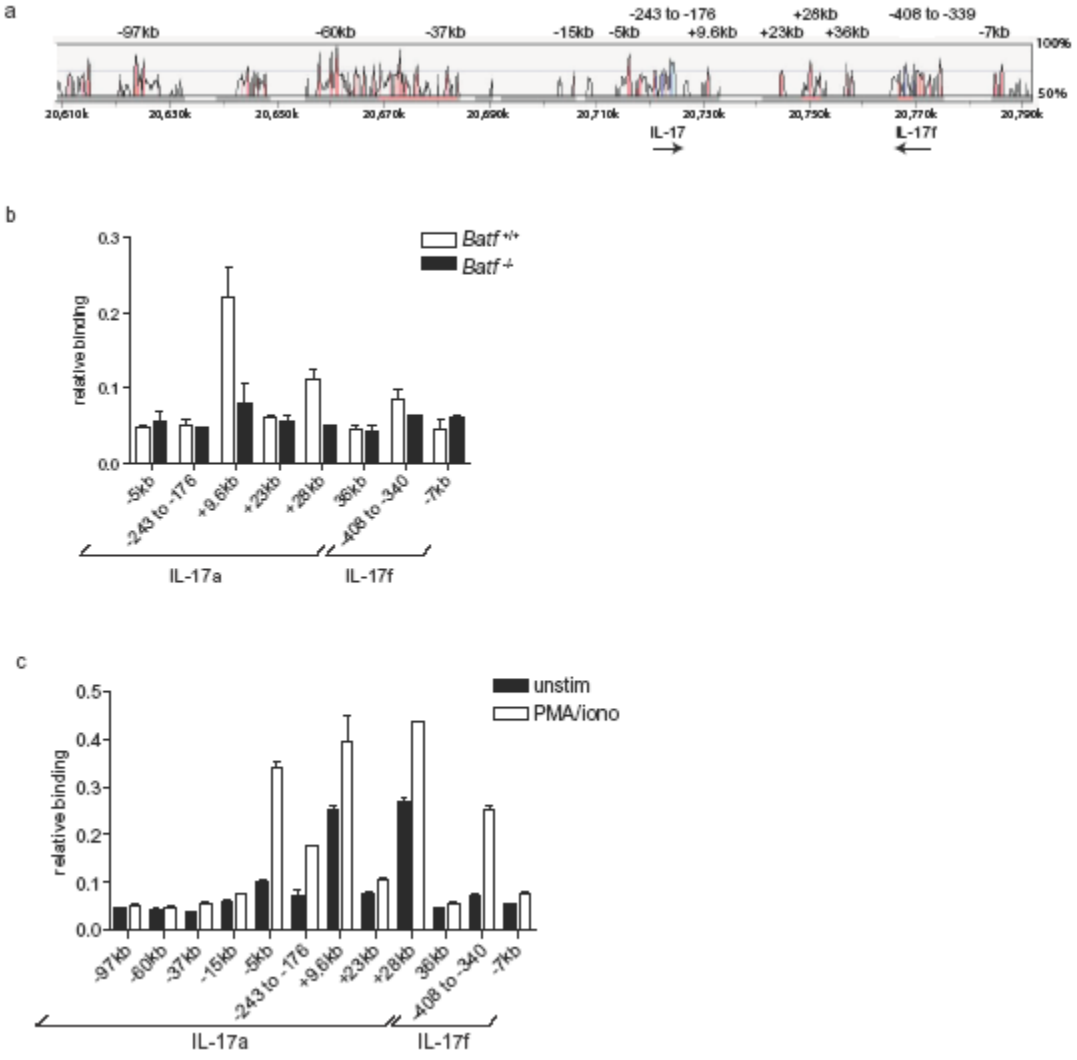
Supplementary Figure A2.S10.



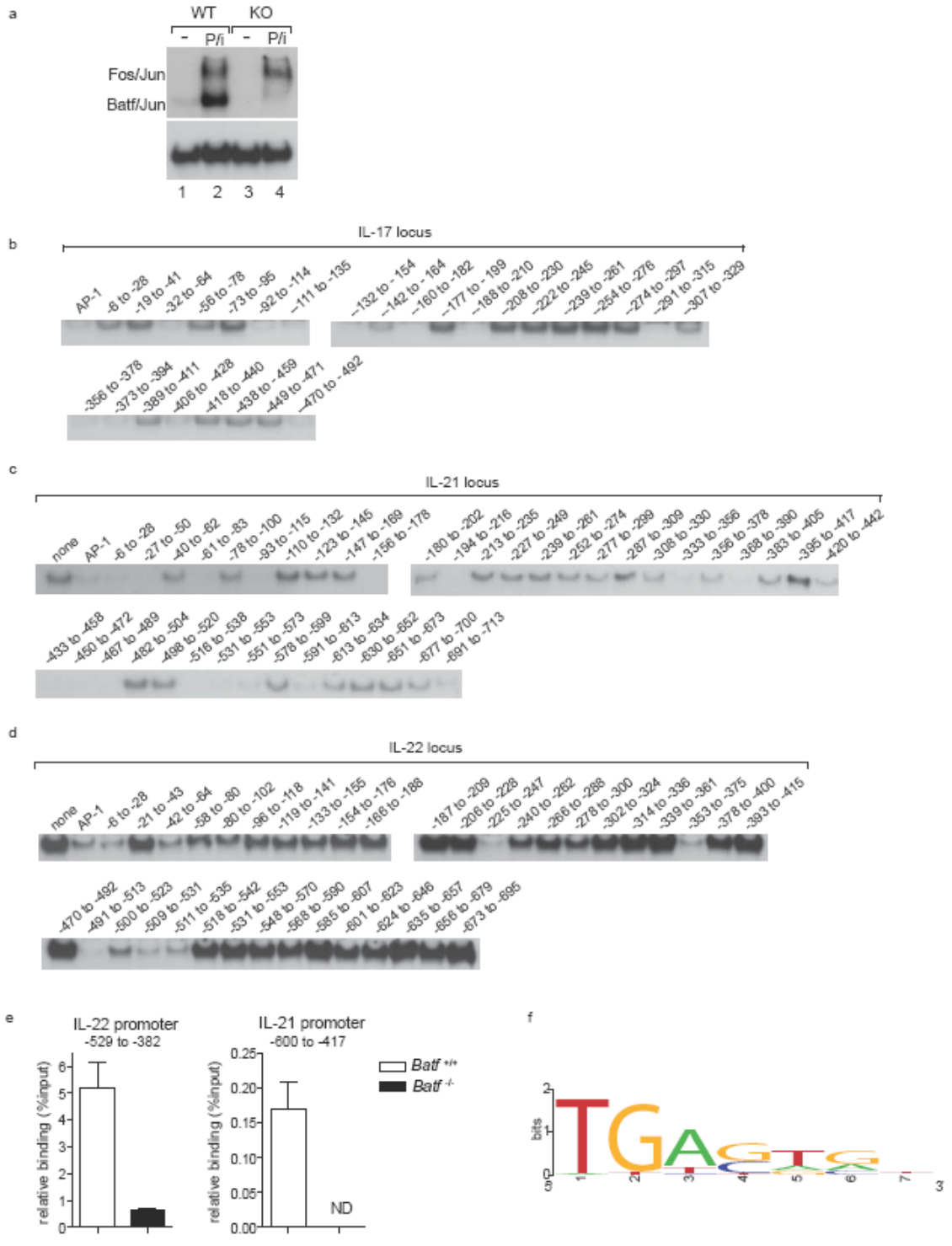
Supplementary Figure A2.S11.



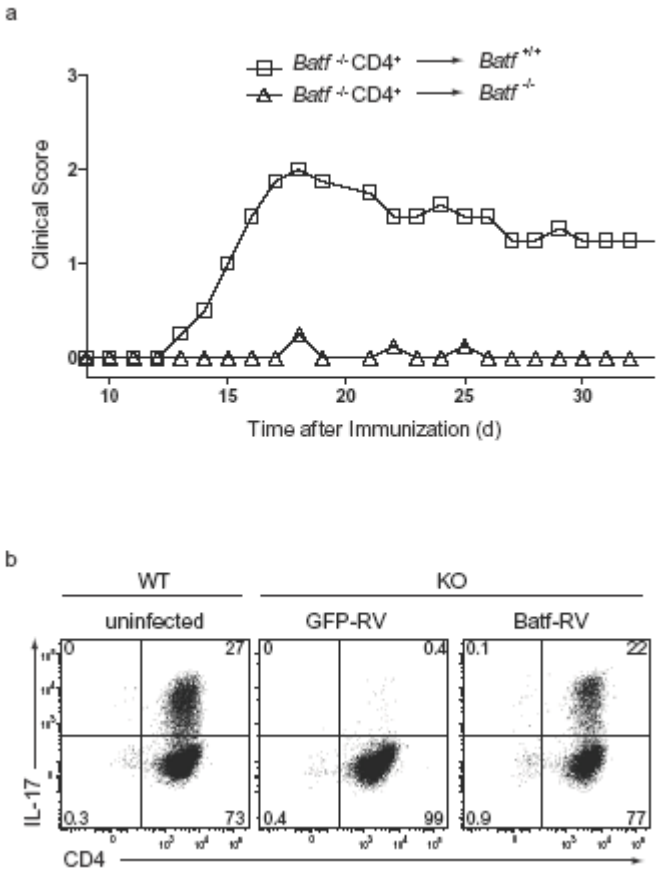
Supplementary Figure A2.S12.



Supplementary Figure A2.S13.



Supplementary Figure A2.S14.



Supplementary Methods

Flow cytometry

All flow cytometric data was collected on a FACS Calibur or FACS Canto (both BD Biosciences) and analyzed using FloJo analysis software (Tree Star, Inc.). The following antibodies were purchased from BD Biosciences; anti-CD4-Allophycoerythrin (APC), CD4-Phycoerythrin (PE)/Cy7 (RM4-5), anti-CD8-APC (53-6.7), anti-CD44-APC (IM7), anti-CD62L-PE (Mel14), anti-CD3-APC (145-2C11), anti-IgM-PE (II/41), anti-B220 Fluorescein isothiocyanate (FITC) (RA3-6B2), anti-IL-17-PE (TC11-18H10), anti-IFNγ-PE (XMG1.2), anti-IFNγ-APC, anti-IL-4-APC (11B11), anti-IL-10-APC (JES5-16E3), anti-CD16/32 (2.4G2), anti-CD11c-PE (HL3), anti-CD11b-PECy7 (M1/70)), anti-CD44-APC (1M7), anti-CD25-APC (3C7), anti-phospho Stat3-AlexaFluor 647 (4/P-Stat3), Streptavidin-PeCy7, 7-AAD, AnnexinV-FITC and AnnexinV staining solution. The following antibodies and solutions were purchased from eBioscience; anti-AA4.1 APC (AA4.1), anti-IgD PE (11-26c), IL-17A-FITC (eBio17B7), anti-Foxp3 (FJK-16s) and Foxp3 staining buffers. Anti-CD4-FITC and anti-CD8-FITC were purchased from Invitrogen. Anti-Dec205-biotin (MG38) was purchased from Cedarlane. CD1d-PBS57-PE and CD1d-unloaded-PE tetramers were obtained from the tetramer facility at the NIH. Anti-IL-22 (RMF 222CK) was purchased from Antigenix.

Isolation of dendritic cells for flow cytometry.

Spleens were isolated, cut into small pieces and digested with Collagenase B (Roche) and DNase I (Sigma) for 30 min at 37^{o}C.

Red blood cells were lysed by incubation with Red Blood Cell Lysis Buffer (Sigma) (1 minute at room temperature). Single cell suspensions were prepared by passing digested spleens through 35μm nylon cell strainers (Fisher Scientific) and were stained with antibodies for analysis by flow cytometry.

Isolation of naïve T cells.

Splenic single cells suspensions were generated and red blood cells were lysed by incubation with Red Blood Cell Lysis Buffer (Sigma) (1 minute at room temperature). Splenocytes were then negatively depleted of $B220^+$ and $CD8^+$ cells using magnetically labeled beads followed by depletion over LD columns (all Miltenyi Biotec). The depleted fraction was then stained with antibodies to CD4, CD62L and CD25 (all BD Biosciences) and $CD4^+CD62L^+CD25^-$cells were sorted on a MoFlo cytometer. Sort purity was generally >98%. For some experiments, as indicated, $CD4^+$ T cells were isolated from spleens by incubation with anti-CD4 magnetic beads and selection via LS columns (Miltenyi Biotec) according to the manufacturer's recommendations.

Cell culture.

For T cell differentiation assays, sorted naïve $CD4^+$ $CD62L^+CD25^-$T cells or magnetically purified $CD4^+$ T cells were isolated as indicated. Cells were cultured at $0.5x10^6$ cells/well in 48 well plates containing plate-bound anti-CD3 (from ascites) and soluble anti-CD28 (37.5; BioXcell; 4μg/ml). Cultures were supplemented with anti-IL-4 (11B11; hybridoma supernatant), IFN-γ (Peprotech; 0.1ng/ml) and IL-12 (Genetics Institute; 10U/ml) for TH1; anti-IFN-γ (H22; BioXcell; 10μg/ml), anti-IL-12 (Tosh; BioXcell; 10 μg/ml) and IL-4 (Peprotech; 10ng/ml) for TH2; anti-IL-4, anti-IL-12, anti-

IFN γ, IL-6 (Peprotech; 20ng/ml) and TGF- β (Peprotech; 0.5ng/ml) for TH17 differentiation. In some experiments, cultures were supplemented with IL-21 (50ng/ml; all Peprotech), anti-IL-6 (MP5-20F3; eBioscience; 10μg/ml), anti-TGF-β (1D11, R&D Biosystems, 10μg/ml) or anti-IL-2 (JES6-1A12; BioXcell; 10μg/ml) as indicated. For TH1 and TH2 conditions and differentiation without addition of cytokines (Supplementary Fig. A2.6a) cells were restimulated on day 7 with anti-CD3 and anti-CD28. Brefeldin A was added for the last 4 hours of stimulation. Unless otherwise indicated cells differentiated under TH17 conditions, were restimulated at the indicated time points with Phorbol 12-myristate 13-acetate (PMA) (50ng/ml; Sigma) and ionomycin (1μM; Sigma) for 4 hours in the presence of Brefeldin A (1μg/ml; Epicentre). Cells were then analyzed by intracellular cytokine staining and flow cytometry. In some experiments, as indicated, magnetically purified $CD4^+$ T cells from DO11.10 transgenic mice were activated with OVA (3μM) and irradiated syngeneic splenocytes in the presence of anti-IL-4, anti-IL-12, anti-IFN γ, IL-6 and TGF- β (1ng/ml) to induce TH17 differentiation. To induce TH17 differentiation in total splenocytes, single cells suspensions from spleens were prepared and red blood cells were lysed. Total splenocytes were activated at $4x10^6$ cells/well in 12 well plates containing plate-bound anti-CD3, anti-IL-4 (hybridoma supernatant), anti-IL-12 (10μg/ml), anti-IFNγ (10μg/ml), IL-6 (20ng/ml) and TGF-β (1ng/ml). Cells were restimulated with PMA and ionomycin for 4h in the presence of Brefeldin A before intracellular cytokine staining and analysis by flow cytometry. For STAT3-phosphorylation assays magnetically purified $CD4^+$ or $CD8^+$ T cells were stimulated with anti-CD3 and anti-CD28 in the presence of IL-6 or IL-21 (50ng/ml) followed by intracellular staining and analysis by flow cytometry.

ELISA.

The concentration of IL-21 in supernatants from $CD4^+$ T cells activated for 3 days under TH17 conditions was determined by ELISA (R&D Systems) according to the manufacturer's recommendations.

Isolation of Lamina Propria T cells.

For isolation of lamina propria T cells, mice were sacrificed; small intestines removed, placed in cold DMEM media (10%FCS) and cleared of Peyer's patches and residual mesenteric fat tissue. Intestines were then opened longitudinally, cleared of contents and cut into 0.5cm pieces. The pieces were washed multiple times in cold media and twice in ice cold Citrate BSA (CB-BSA) buffer followed by two 15 minute incubations in CB-BSA with agitation. After each incubation cells were vortexed to remove epithelial cells. The remaining intestinal pieces were then washed twice with cold media before digestion in media containing 75U/ml Collagenase IV (Sigma) at 37°C for 1 hour. The solution was vortexed at 20 min intervals to detach lymphocytes. After one hour the solution was filtered through a 35μm strainer, the pieces were collected and digested a second time. Supernatants from both digestions were combined, washed once, suspended in the 70% fraction of a percoll gradient and overlaid with 37% and 30% percoll gradient fractions. Lymphocytes were collected at the 70-37% interface, washed once in PBS and stimulated with PMA/ionomycin in the presence of Brefeldin A for 3 hours before cells were stained for extracellular markers and intracellular cytokines.

Induction of EAE and disease scoring.

Age and sex matched mice (7-10 weeks old) were immunized subcutaneously with 100

μg MOG35-55 peptide (Sigma) emulsified in
CFA (IFA supplemented with 500μg *Mycobacterium tuberculosis*) on day 0. On days 1 and 3, mice were injected with 300ng Pertussis Toxin (List Biological Laboratories) intraperitonally (i.p.). Clinical scores were given on a scale of 1-5 as follows: 0, no overt signs of disease; 1, limp tail or hind limb weakness, but not both; 2, limp tail and hind limb weakness; 3, partial hind limb paralysis; 4, complete hind limb paralysis; 5, moribund state or death by EAE. Mice with a score of 4 were given 300 μl saline solution subcutaneously to prevent dehydration. Mice with a score of 5 were euthanized. Some mice died during the course of the experiment. Their clinical score of 5 was included in the analysis for the remainder of the experiment. For T cell transfer experiments, $CD4^+$ T cells were isolated from splenic single cell suspensions by magnetic separation with anti-CD4 magnetic beads and positive selection via LS columns (Miltenyi Biotec). $1x10^7$ MACS purified $CD4^+$ T cells were injected i.p. on day -4 followed by EAE induction on day 0 as described above.

Isolation of CNS lymphocytes.

Brain and spinal cords were removed from mice after perfusion with 30ml of saline solution. Single cell suspensions were prepared by dispersion through sterile 35μ nylon cell strainers (Fisher Scientific) and mixed at room temperature for 1h in HBSS containing 0.1% collagenase, 0.1μg/ml TLCK (N-α-tosyl-L-lysine chloromethylketone hydrochloride), and 10μg/ml DNaseI (all Sigma). The resulting suspension was pelleted, resuspended in the 70% fraction of a Percoll gradient
and overlaid by additional 37% and 30% layers. The Percoll gradient separation was achieved by centrifugation for 20 min at 2000rpm and lymphocytes were collected at the

70-37% interface. Subsequently cells were activated with PMA and ionomycin for 3-4 hours in the presence of Brefeldin A and intracellular cytokine staining was performed.

Real time PCR.

Naïve $CD4^+CD62L^+CD25^-$T cells were isolated by cell sorting and activated with plate-bound anti-CD3 and soluble anti-CD28 antibodies under TH17 conditions for 3 days, unless otherwise indicated. Total RNA was isolated from the indicated cells using Quiagen RNeasy Mini Kit and cDNA was synthesized using SuperscriptIII reverse transcriptase (Invitrogen). Real time PCR analysis was performed using ABI SYBR Green master mix according to the manufacturer's instructions on an ABI7000 machine (Applied Biosystems) using the relative standard curve method. The PCR conditions were 2min at 50°C, 10 min at 95°C followed by 40 2-step cycles of 15s at 95°C and 1min at 60°C. Primers for RORγt (RORγt forward 5'-CGCTGAGAGGGCTTCAC, RORγt reverse 5'-GCAGGAGTAGGCCACATTACA)[39], IL-21 (IL-21 forward 5'-ATCCTGAACTTCTATCAGCTCCAC, IL-21 reverse 5′-GCATTTAGCTATGTGCTTCTGTTTC)[40], IL-22 (IL-22 forward-5' CATGCAGGAGGTGGTACCTT, IL-22 reverse- 5'-CAGACGCAAGCATTTCTCAG)[41], RORα (RORα forward 5'-TCTCCCTGCGCTCTCCGCAC, RORα reverse 5'-TCCACAGATCTTGCATGGA)[38], IRF-4 (IRF-4 forward 5'-GCCCAACAAGCTAGAAAG, IRF-4 reverse: 5'-TCTCTGAGGGTCTGGAAACT) [42] and HPRT as normalization control (HPRT forward 5'-AGCCTAAGATGAGCGCC, HPRT reverse 5'-TTACTAGGCAGATGGCCACA) were used to evaluate relative gene expression. For analysis of acute phase response proteins, mice were injected intraperitonally with either 0.9% saline solution or IL-6

(0.3μg per mouse) in 0.9% saline solution. Four hours later, total liver RNA was isolated using Trizol reagent (Invitrogen) according to the manufacturer's recommendations. cDNA was synthesized and real time PCR performed as described above. Primers used for serum amyloid protein P (SAP forward: 5'-TTTCAGAAGCCTTTTGTCAGA and SAP reverse: 5'-AAGGTCACTGTAGGTTCGGA) [43], c-reactive protein (CRP forward: 5'-TTCTGGATTGATGGGAAAAGC and CRP reverse: 5'-AAACATTGGGGCTGAGTGTC)[43], Serum amyloid protein A (SAA forward 5'-TCTCTGGGGCAACATAGTATACCTCTCAT and SAA reverse 5'-TTTATTACCCTCTCCTCCTCAAGCAGTTAC) [44], fibrinogen β (fibß forward: 5'-ATTAGCCAGCTTACCAGGATGGGACCCAC-3', Fibß reverse: 5'-CAGTAGTAT CTGCCGTTTGGATTGGCTGC-3')[45], alpha-1-acid glycoprotein (AGP forward: TCT CTG AAC TCC GAG GGC TG AGP reverse: GAGACAGAATCAAAGTGCACAGGA)[46] and HPRT as normalization control (HPRT forward 5'-AGCCTAAGATGAGCGCC, HPRT reverse 5'-TTACTAGGCAGATGGCCACA) were used to evaluate relative gene expression.

Gene expression profiling.

Naïve $CD4^+$ $CD62L^+$ $CD25^-$T cells and $CD4^+$ $CD62L^+$ $CD25^+$ regulatory T cells were isolated from C57BL/6 mice. Naïve $CD4^+$ $CD62L^+$ $CD25^-$T cells were differentiated under TH1 and TH2 conditions for 7 days. After restimulation with anti-CD3 and anti-CD28 for 24h, TH1 and TH2 cells were sorted for IFN-γ and IL-4 production respectively using cytokine secretion assays (Miltenyi Biotec) according to the manufacturer's recommendations. For gene expression profiling of TH17 cells, naïve $CD4^+$ $CD62L^+$ $CD25^-$ T cells were activated for 3 days with anti-CD3 and anti-CD28 in

the presence of anti-IL-4, anti-IL-12, anti-IFNγ, anti-IL-2, IL-6 and TGF-β (0.5ng/ml). For gene expression analysis in *Batf*$^{-/-}$T cells, naive CD4^{+} CD62L^{+} CD25^{-} T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were activated for 3 days with anti-CD3 and anti-CD28 in the presence of either anti-IL-4, anti-IL-12, anti-IFNγ, IL-6 and TGF-β (0.5ng/ml); anti-IL-4, anti-IL-12, anti-IFNγ, IL-6 and anti-TGF-β; anti-IL-4, anti-IL-12, anti-IFNγ, anti-IL-6 and TGF-β or anti-IL-4, anti-IL-12, anti-IFNγ, anti-IL-6 and anti-TGF-β. IL-2 was neutralized in all conditions. Total RNA was isolated from cells using Quiagen RNeasy Mini Kit. Biotinylated antisense cRNA was generated using two cycle target preparation kit (Affymetrix). After fragmentation, cRNA was hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 Arrays. Data were normalized and expression values were modeled using DNA-Chip analyzer (dChip) software (www.dChip.org).

Retroviral infection and analysis.

mRNA was isolated from 129SvEv total thymocytes using Quiagen RNAeasy Mini Kit and cDNA was amplified by SuperscriptIII (Invitrogen). Murine RORγt transcript was amplified using primers 5'-CTCGAGGTGTGCTGTCCTGGGCTAC and 5'-CTCGAGGGGAGACGGGTCAGAGGG. Underlined nucleotides indicate XhoI overhangs used to clone RORγt into XhoI digested GFP-RV retrovirus[47] or XhoI digested hCD4-RV[48]. *Batf* cDNA was cloned from CD4^{+} T cell mRNA using primers 5'-GGAAGATTAGAACCATGCCTC and 5'-AGAAGGTCAGGGCTGGAAG and subcloned into the GFP-RV retrovirus[47]. An N-terminal FLAG tag was introduced by Quick Change Mutagenesis kit (Stratagene) using the primers 5'-**G**GACTACAAAGACGATGACGACAAGCCTCACAGCTCCGACAGCA and 5'-CTTGTCGTCATCGTCTTTGTAGTCCATGGTTCTAATCTTCCAGATC. The

underlined sequence indicates nucleotides used to introduce the FLAG-tag. The retrovirus based reporter hCD4-pA-GFP-RV[48], in which a cytoplasmic truncated human CD4 (hCD4) marks viral infection and green fluorescence protein (GFP) is used to report promoter activity has been described previously and was modified as follows to generate hCD4-pA-GFP-RV-IL-17p. The 1021bp promoter region of murine IL-17a was generated by PCR from genomic 129SvEv DNA using primers 5'-AGCTTGAACAGGAGCTATCGGTCC and 5'-AAGCTTGAGGTGGATGAAGAGTAGTGC. Underlined nucleotides indicate overhangs containing HindIII restriction sites used to clone the resulting PCR product into hCD4-pA-GFP-RV. Retroviral vectors were packaged in Phoenix E cells as described previously[47]. Magnetically purified $CD4^+$ T cells were infected with viral supernatants on days 1 and 2 after activation with anti-CD3 and anti-CD28. 3 days after activation cells were restimulated with PMA/ionomycin in the presence of Brefeldin A and analyzed by intracellular cytokine staining and Flow Cytometry. For the experiments in Figure 4, $CD4^+$ T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ mice were activated under TH17 conditions and infected with the IL-17 reporter virus. Stably infected T cells were restimulated with PMA/ionomycin for 4h and examined for GFP expression on day 3 after initial activation.

Statistical Analysis.

A Student's unpaired two-tailed t-test was used to indicate statistically significant differences between indicated groups. Differences with a P value <0.05 were considered significant.

Electrophoretic mobility shift assays.

Whole cell extracts were prepared from total splenocytes activated for 3 days with anti-CD3, TGF-β and IL-6 as described previously[49]. For EMSA analysis the AP-1 consensus probe[36] (top: AGCTTCGCTTGATGAGTC and bottom: GCCGACTGAGTAGTTCGC), RORE element in CNS2 of the IL-17 gene[38] (top: GAAAGTTTTCTGACCCACTTTAAATCA and bottom: CTTTAACTAAATTTCACCCAGTCTTTT) and -187 to -155 of the IL-17 promoter (top: GGTTCTGTGCTGACCTCATTTGAGGATG and bottom: AAAAGACTGGGTGAAATTTAGTTAAAG), Eα Y box probe (TCGACATTTTTCTGATTGGTTAAAAGTC)[50] were used after labeling with ^{32}P-dCTP. The probe ($2.5x10^4$cpm per reaction) was used along with 15μg of total cell extracts and 1ug poly diDC as described previously[50]. For competitor-supershift assay, Batf binding to the AP-1 consensus probe[36] was assessed by anti-FLAG supershift. Unlabeled probes from the IL-17a, IL-21 and IL-22 promoters (Supplementary table A2.S6) were used to compete for Batf binding to the AP-1 consensus probe. Single stranded overhangs of the competitor oligos were not filled in. Sequences identified as competitors for Batf binding were used to determine the Batf consensus motif. For supershift analysis of the EMSA complexes formed on the AP-1 probe, whole cell extracts were prepared as above. 8μg whole cell extracts were incubated for 15min on ice with anti-Batf, anti-Fos (K25), anti-c-Jun (D), anti-c-Jun (N), anti-JunB (C11), anti-JunD (329), anti-ATF-1 (H60) and anti-ATF-3 (C-19) (all Santa Cruz Biotechnology) before $2.5x10^4$cpm of the AP-1 consensus probe was added. To test whether Batf binding to the AP-1 probe requires stimulation DO11.10 transgenic $CD4^+$ T cells were activated for 3 days with OVA, irradiated APCs,

anti-IFN-γ/IL-4/IL12, TGF-β and IL-6, followed by a period of 3 days rest in the presence of TGF-β and IL-6. Cells were left untreated or activated with PMA/ionomycin for 4 hrs before whole cell extracts were prepared and used in EMSA analysis as described above. Reactions were electrophoresed on 7.5 or 10% bisacrylamide gels to achieve optimal band separation.

CONSENSUS program for determination of Batf binding motif.

Sequences of the proximal promoter regions of IL-17, IL-21, and IL-22 identified as competitors for Batf binding in the competitor-supershift EMSA assay were input into CONSENSUS version v6d[51]. Default program parameters were applied, except for searching the reverse complement of the input sequences (c2) and uniform background nucleotide frequencies. The program was searching potential motif lengths from 5 to 15 using the expected frequency statistic (e-value) and the optimal motif length was determined as 7. The corresponding weight matrix, with a sample size adjusted information content of 4.467, was chosen from the final cycle. The enrichment of the binding motif in the input set was verified using PATSER v3e[52]. Using the numerically calculated cutoff score, 38/40 of the input training sequences were identified as containing the motif. The motif is presented as a Weblogo[37] in which the size of each nucleotide is proportional to the frequency of its appearance at each position.

Batf Chromatin immunoprecipitation (ChIP).

ChIP was performed as previously described[53] using an affinity purified anti-Batf rabbit polyclonal antibody prepared by Brookwood Biomedical (Birmingham, AL). Briefly, chromatin was prepared from $1x10^7$ $CD4^+$ T cells isolated from C57BL/6 *Batf*$^{+/+}$ mice

stimulated under TH17 polarizing conditions with anti-CD3 (2.5μg/ml) and syngeneic splenic feeder cells, then restimulated or not at the indicated time points with PMA (50ng/ml) and ionomycin (750ng/ml) for 4 h. For experiments assessing early binding of Batf to the DNA CD4$^+$ T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ 129SvEv mice were activated with anti-CD3/CD28 coated beads under TH17 conditions for 24 hours, then processed for ChIP analysis. Immunoprecipitations were performed with 20 μg/ml anti-Batf rabbit polyclonal antibody using the Chromatin Immunoprecipitation (ChIP) Assay Kit from Millipore (Billerica, MA) according to the manufacturer's recommendations. Immunoprecipitated DNA released from cross-linked proteins was quantitated by real-time PCR as previously reported[53], and was normalized to input DNA. Unless otherwise indicated data are presented as mean + s.d from 2 independent experiments. All real-time PCR primers and probes are included in Supplementary table A2.S5. The analyzed sites are denoted relative to the ATG start codons for the *Il17a* or *Il17f* gene. For ChIP analysis of the IL-21 and IL-22 promoters DO11.10 transgenic CD4$^+$ T cells from *Batf*$^{+/+}$ and *Batf*$^{-/-}$ were stimulated with OVA and APC for 3 days, rested for 3 days before restimulation with PMA/ionomycin for 4h on day 5 and processing for ChIP as described above. Real time PCR analysis was performed using ABI SYBR Green master mix according to the manufacturer's instructions on a Step One Plus (Applied Biosystems) using the relative standard curve method. Results were normalized to input DNA. Sequences of primers used in the analysis are given in Supplementary Table A2.S5.

Supplementary References

35. Sun,Z. *et al.* Requirement for RORgamma in thymocyte survival and lymphoid organ development. *Science* **288**, 2369-2373 (2000).

36. Echlin,D.R., Tae,H.J., Mitin,N. & Taparowsky,E.J. B-ATF functions as a negative regulator of AP-1 mediated transcription and blocks cellular transformation by Ras and Fos. *Oncogene* **19**, 1752-1763 (2000).

37. Crooks,G.E., Hon,G., Chandonia,J.M. & Brenner,S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

38. Yang,X.O. *et al.* T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. *Immunity* **28**, 29-39 (2008).

39. Ivanov,I.I. *et al.* The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells. *Cell* **126**, 1121-1133 (2006).

40. Zhou,L. *et al.* IL-6 programs T(H)-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nat Immunol* **8**, 967-974 (2007).

41. Chung,Y. *et al.* Expression and regulation of IL-22 in the IL-17-producing CD4+ T lymphocytes. *Cell Res* **16**, 902-907 (2006).

42. Negishi,H. *et al.* Negative regulation of Toll-like-receptor signaling by IRF-4. *Proc. Natl. Acad. Sci U. S. A* **102**, 15989-15994 (2005).

43. Korbelik,M., Cecic,I., Merchant,S. & Sun,J. Acute phase response induction by cancer treatment with photodynamic therapy. *Int. J Cancer* **122**, 1411-1417 (2008).

44. Dierssen,U. *et al.* Molecular dissection of gp130-dependent pathways in hepatocytes during liver regeneration. *J Biol Chem.* **283**, 9886-9895 (2008).

45. Chauvet,C. *et al.* The gene encoding fibrinogen-beta is a target for retinoic acid receptor-related orphan receptor alpha. *Mol Endocrinol.* **19**, 2517-2526 (2005).

46. Theilgaard-Monch,K. *et al.* Highly glycosylated alpha1-acid glycoprotein is

synthesized in myelocytes, stored in secondary granules, and released by activated neutrophils. *J Leukoc. Biol* **78**, 462-470 (2005).

47. Ranganath,S. *et al.* GATA-3-dependent enhancer activity in IL-4 gene regulation. *J. Immunol.* **161**, 3822-3826 (1998).

48. Zhu,H. *et al.* Unexpected characteristics of the IFN-gamma reporters in nontransformed T cells. *J. Immunol.* **167**, 855-865 (2001).

49. Nakshatri,H. & Currie,R.A. Differential whole-cell extract preparation and electrophoretic mobility shift assay to evaluate the effect of tyrosine phosphatases on DNA binding activity of transcription factors. *Anal. Biochem* **236**, 178-181 (1996).

50. Szabo,S.J., Gold,J.S., Murphy,T.L. & Murphy,K.M. Identification of cis-acting regulatory elements controlling interleukin-4 gene expression in T cells: roles for NF-Y and NF-ATc [published erratum appears in Mol Cell Biol 1993 Sep;13(9):5928]. *Mol. Cell Biol.* **13**, 4793-4805 (1993).

51. Hertz,G.Z. & Stormo,G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* **15**, 563-577 (1999).

52. Stormo,G.D., Schneider,T.D., Gold,L. & Ehrenfeucht,A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997-3011 (1982).

53. Hatton,R.D. *et al.* A Distal Conserved Sequence Element Controls Ifng Gene Expression by T Cells and NK Cells. *Immunity* (2006).

Curriculum vitae

Gurmukh Singh Sahota, MS

Contact Information

Gurmukh Singh Sahota

Washington University

Department of Genetics

4444 Forest Park Parkway

Campus Box 8510, Room 5401

St. Louis, MO 63108

Education

2012	MD/PhD	Computational and Systems Biology, Washington University in St. Louis
2004	MS	Biophysics/Computational Biology, University of Illinois at UC
2002	BS/BA	Computer Science/Molecular Biology and Biochemistry, Rutgers University

Professional Experience

2005-2010	Graduate Research Assistant, Washington University, Professor Gary Stormo
2003-2004	Princeton Review MCAT instructor.
2002-2005	Graduate Research Assistant, University of Illinois at UC, Professor Eric Oldfield
1999-2002	Undergraduate Research Assistant, Rutgers University, Professor Gaetano Montelione

Honors and Awards

2007 Selected by Deans for professionalism study of medical students (Washington University)

2003 Royalties received for software development and licensing (Rutgers University)

2002 *Summa cum laude* (highest honors) and Henry Rutgers Honors (Rutgers University)

Grants and Scholarships

2007-2009 NIH Computational Biology Training Grant (Washington University)

2005-2007 NIH MSTP Training Grant (Washington University)

2002-2004 NIH Molecular Biophysics Training Grant (University of Illinois at UC)

2001, 2002 Travel Awards for Experimental Nuclear Conference (ENC)

2001 David and Dorothy Bernstein Summer Research Award (Rutgers University)

1998-2002 Outstanding Student Recruitment Award (Full Tuition Scholarship) (Rutgers University)

1998-2002 Edward J. Bloustein Scholarship (Rutgers University)

Publications

1. Baran, M. C., Moseley, H. N., **Sahota, G.** & Montelione, G. T. SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. *J Biomol NMR* **24**, 113-21 (2002).
2. Moseley, H. N., **Sahota, G.** & Montelione, G. T. Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J Biomol NMR* **28**, 341-55 (2004).
3. Franks, W. T., Zhou, D. H., Wylie, B. J., Money, B. G., Graesser, D. T., Frericks, H. L., **Sahota, G.** & Rienstra, C. M. Magic-angle spinning solid-state NMR spectroscopy of the beta1 immunoglobulin binding domain of protein G (GB1): 15N and 13C chemical shift assignments and conformational analysis. *J Am Chem Soc* **127**, 12291-305 (2005).
4. Ling, Y., **Sahota, G.**, Odeh, S., Chan, J. M., Araujo, F. G., Moreno, S. N. & Oldfield, E.

Bisphosphonate inhibitors of Toxoplasma gondi growth: in vitro, QSAR, and in vivo investigations. *J Med Chem* **48**, 3130-40 (2005).

5. Kotsikorou, E., **Sahota, G.** & Oldfield, E. Bisphosphonate inhibition of phosphoglycerate kinase: quantitative structure-activity relationship and pharmacophore modeling investigation. *J Med Chem* **49**, 6692-703 (2006).
6. Schaaf, C. A., Misulovin, Z., **Sahota, G.**, Siddiqui, A. M., Schwartz, Y. B., Kahn, T. G., Pirrotta, V., Gause, M. & Dorsett, D. Regulation of the Drosophila Enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins. *PLoS One* **4**, e6202 (2009).
7. Schraml, B. U., Hildner, K., Ise, W., Lee, W. L., Smith, W. A., Solomon, B., **Sahota, G.**, Sim, J., Mukasa, R., Cemerski, S., Hatton, R. D., Stormo, G. D., Weaver, C. T., Russell, J. H., Murphy, T. L. & Murphy, K. M. The AP-1 transcription factor Batf controls T(H)17 differentiation. *Nature* **460**, 405-9 (2009).
8. **Sahota, G.** & Stormo, G. D. Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. *Bioinformatics* **26**, 2672-7 (2010).